INVESTIGACION Y CIENCIA

Edición en español de SCIENTIFIC AMERICAN



AERONAVES DE PROPULSION HUMANA

Enero 1986
450 Ptas.
(IVA incluido)

Los espacios en gris
corresponden a publicidad
en la edición impresa

Número 112
Enero 1986

INVESTIGACION Y
CIENCIA
Edición en español de SCIENTIFIC AMERICAN

INVESTIGACION Y CIENCIA

DIRECTOR Francisco Gracia Guillén
REDACCION José María Valderas Gallardo (Redactor Jefe)
Carlos Oppenheimer
José María Farré Josa
PRODUCCION César Redondo Zayas
PROMOCION EXTERIOR Pedro Clotas Cierco
EDITA Prensa Científica, S.A.
Calabria, 235-239
08029 Barcelona (ESPAÑA)

Colaboradores de este número:

Asesoramiento y traducción:

Isabel García Acha y Julio Rodríguez Villanueva: *Terapéutica y lucha contra el cáncer*; Ramón Pascual: *Rayos cósmicos de Cisne X-3*; Anunciación Ilundáin: *La señal del calcio*; Montserrat Domingo: *Retazos litosféricos*; Luis Palacios: *Respiración cutánea en los vertebrados*; Donato Franco: *Vuelo de propulsión humana*; Luis Bou: *Tarjetas inteligentes y Juegos de Ordenador*; J. Vilardell: *Taller y laboratorio.*

Ciencia y sociedad:

Pedro Miguel Echenique, J. R. Germà Lluch y Ramón Pascual

Libros:

Javier de Lorenzo, Ramón Margalef, Salvador Reguant y Jorge Domingo



PORTADA

La ilustración de la portada muestra la aeronave de propulsión humana *Monarch B* en vuelo (véase "Vuelo de propulsión humana", por Mark Drela y John S. Langford, en este número). El *Monarch B* fue diseñado y construido en el Instituto de Tecnología de Massachusetts. El piloto impulsa la aeronave pedaleando una cadena que mueve la hélice. Con 0,5 caballos de vapor, alcanza velocidades de 30 a 37 kilómetros por hora. Para proporcionar potencia adicional, necesaria para elevarse, el piloto enciende un motor eléctrico impulsado por una batería que ha cargado antes del vuelo pedaleando un generador. Las manos del piloto empuñan una palanca de control por medio de la cual puede hacer que la nave suba, baje o gire. El *Monarch B* tiene una envergadura de 18,60 metros y 12.200 metros. A velocidad de crucero la hélice gira a unas 210 rpm.

Suscripciones:

Prensa Científica, S. A.
Calabria, 235-239
08029 Barcelona (España)
Teléfono 322 05 51 ext. 33-37

Condiciones de suscripción:

España:
Un año (12 números):
4400 pesetas (IVA incluido)

Extranjero:
Un año (12 números): 33 U.S. $

Ejemplar atrasado ordinario:
450 pesetas (IVA incluido)

Ejemplar atrasado extraordinario:
575 pesetas (IVA incluido)

Distribución para España:

Distribuciones de Enlace, S. A.
Bruch, 49 - 08009 Barcelona

Distribución para los restantes países:

Editorial Labor, S. A.
Calabria, 235-239 - 08029 Barcelona

Publicidad:

Madrid:
Gustavo Martínez Ovin
Avda. de Moratalaz, 137 - 28030 Madrid
Teléfonos 430 84 81 - 437 12 91

Cataluña:
Paulino Muñoz Victoria
Comunicación Diaria, S.A.
Aribau, 195, 4.° E - 08021 Barcelona
Teléfono 200 17 58

Controlado por O.J.D.

PROCEDENCIA DE LAS ILUSTRACIONES

Pintura de la portada de Ted Lodigensky

Página	Fuente	Página	Fuente
7-14	Ilil Arbel	40	J. R. Germà Lluch y Ricardo Génova
17	Ian Worpole	43	EROS Data Center
18-20	Andrew Christie	44-45	Benita L. Murchey y David L. Jones, Servicio Geológico de los Estados Unidos
21	Trevor C. Weekes, Observatorio Fred Lawrence Whipple (*arriba*); Andrew Christie (*abajo*)	46	Anita G. Harris, Servicio Geológico de los Estados Unidos
22	Rick F. Harnden, Jr., Centro de Astrofísica del Observatorio del Harvard College y del Observatorio Smithonsiano de Astrofísica	47-52	Hank Iken, Walken Graphics
		55	Enid Kotschnig
23	Trevor C. Weekes, Observatorio Fred Lawrence Whipple, y John C. Geary, Centro de Astrofísica del Observatorio del Harvard College y del Observatorio Smithsoniano de Astrofísica	56-59	Tom Prentiss
		60	Warren W. Burggren y Alan W. Pinder, Universidad de Massachusetts en Amherst
		61-62	Tom Prentiss
		64-65	Steven Finberg
		66-71	Hank Iken, Walken Graphics
24-25	Andrew Christie	73	James Kilkelly
29	David Bacon, Universidad de Alberta	74-80	Jerome Kuhl
		83	Agencia Espacial Europea
30-34	George V. Kelvin, Science Graphics	84-92	J. Meseguer, A. Sanz y Joan Cotoner
35	Feroze N. Ghadially, Universidad de Saskatchewan	94-98	Ilil Arbel
		101-102	Observatorios de Monte Wilson y Las Campanas
36-37	George V. Kelvin, Science Graphics	103-107	Michael Goodman

ISSN 0210-136X
Dep. legal: B. 38.999-76
Fotocomposición Tecfa, S.A.
Pedro IV, 160 - 08005 Barcelona
Fotocromos reproducidos por GINSA, S.A.
Gráfica Elzeviriana, S.A.
Nápoles, 249 - Tel. 207 40 11
08013 Barcelona
Printed in Spain - Impreso en España

Los autores

JOHN CAIRNS ("Terapéutica y lucha contra el cáncer") es profesor de microbiología de la Facultad de Salud Pública de la Universidad de Harvard. Tras licenciarse en medicina por la Universidad de Oxford marchó a Australia para estudiar la multiplicación de los virus y la estructura de las moléculas de ADN. Se trasladó luego a Estados Unidos y, hasta 1968, dirigió el laboratorio de biología cuantitativa de Cold Spring Harbor. En Harvard, Cairns divide su atención profesional entre el estudio de problemas de salud pública y la investigación sobre el origen de las mutaciones.

P. KEVIN MACKEOW y TREVOR C. WEEKES ("Rayos cósmicos de Cisne X-3") han colaborado en proyectos de investigación en física desde antes de graduarse. MacKeown enseña esa disciplina en la Universidad de Hong Kong. Aunque comenzó su formación en el University College de Dublín se doctoró por la Universidad de Durham. Ha desarrollado cursos de investigación en el Instituto Tata de Investigación Fundamental de Bombay, en la Universidad de Maryland en College Park y en la estatal de Louisiana antes de trasladarse a Hong Kong en 1970. En su trabajo MacKeown ha ido literalmente de un extremo a otro: estaciones de rayos cósmicos en los Andes y experimentos de neutrinos en una mina de oro. Weekes trabaja, desde 1967, como astrofísico en el Observatorio Fred Lawrence Whipple, en Arizona.

ERNESTO CARAFOLI y JOHN T. PENNISTON ("La señal del calcio") dan clases de bioquímica en el Instituto Politécnico Suizo y en la Facultad de Medicina de la Universidad Mayo, respectivamente. Carafoli es italiano de origen y formación, y procede de la Universidad de Módena. Estuvo algunos años en la Johns Hopkins en calidad de asociado postdoctoral. En 1973 alcanzó la plaza que hoy ocupa. Penniston se doctoró en la Universidad de Harvard en 1962, con una tesis sobre ácidos nucleicos. Tras un año de experiencia docente en Pomona College, enseñando química, se trasladó al Instituto de Investigación Enzimática de la Universidad de Wisconsin en Madison, en donde se interesó por las membranas celulares. Desde 1971 hasta 1976 fue profesor asociado de química de la Universidad de Carolina del Norte en Chapel Hill. En 1976 aceptó el cargo de profesor de bioquímica de la Facultad de Medicina de la Universidad Mayo.

DAVID G. HOWELL ("Retazos litosféricos") es investigador del Servicio Geológico de los Estados Unidos. Se licenció en la Universidad Colgate en 1962 y realizó estudios de grado en la Universidad de California en Santa Bárbara, donde se licenció en 1969. Sirvió tres años en el ejército norteamericano antes de obtener su doctorado en Santa Bárbara. Luego se incorporó al Servicio Geológico. Desde 1980 pertenece al claustro docente de la Universidad de Stanford.

MARTIN E. FEDER y WARREN W. BURGGREN ("Respiración cutánea en los vertebrados") son respectivamente profesor adjunto del departamento de anatomía de la Universidad de Chicago y profesor agregado del departamento de zoología de la Universidad de Massachusetts en Amherst. Feder se graduó por la Universidad de Cornell en 1973 y recibió el grado de doctor en ciencias por la de California en Berkeley en 1977. Pasó dos años como becario postdoctoral en la Universidad de Chicago y un año en calidad de profesor e investigador invitado en la Universidad de Silliman en Filipinas, antes de ocupar la plaza actual. Burggren cursó su carrera en la Universidad de Calgary y sus estudios como postgraduado en la de East Anglia. Después de dos años como becario postdoctoral en la Universidad de Britsh Columbia entró en la de Massachusetts. En 1982 Burggren fue nombrado profesor agregado.

MARK DRELA y JOHN S. LANGFORD ("Vuelo de propulsión humana") preparan la defensa de su tesis doctoral en el Departamento de Aeronáutica y Astronáutica del Instituto de Tecnología de Massachusetts y han colaborado en dos proyectos de aeronave de propulsión humana. Ambos cursaron su carrera de ingeniero aeronáutico también en el MIT. Langford es, además, licenciado en ciencias políticas; trabaja en el Instituto de Análisis para la Defensa, de Alejandría (Virginia), en un campo donde convergen política y aeronáutica. Los dos están ilusionados con su proyecto *Dédalo*, un estudio de la posibilidad de un vuelo de propulsión humana entre Creta y Grecia continental.

ROBERT McIVOR ("Tarjetas inteligentes") es director de fabricación de la Motorola Inc., de Austin, Texas. Se formó en el Kirkaldy Technical College, de donde pasó a la Universidad de Strathclyde, en su Escocia natal. Ingresó en Motorola en 1970, y tras ascender hasta el puesto de director de operaciones en Escocia se trasladó a los Estados Unidos. Se cuentan entre sus responsabilidad todas las operaciones de fabricación y control de producción de microprocesadores, así como la dirección de proyectos para clientes importantes. McIvor desempeña un papel de primera línea en la coordinación y gestión de la participación de Motorola en el desarrollo de tarjetas inteligentes, y forma parte del comité nacional que está en vías de establecer una normalización uniforme para estos dispositivos.

JOSE MESEGUER RUIZ y ANGEL SANZ ANDRES ("Comportamiento de los líquidos en ingravidez") son profesores del departamento de aerodinámica de la Escuela Técnica Superior de Ingenieros Aeronáuticos de la Universidad Politécnica de Madrid. Meseguer se incorporó a dicho departamento en 1974, trabajando desde entonces en diversos contratos de la Agencia Europea del Espacio (ESA) relacionados con el control térmico de naves espaciales. Desde 1977 participa también en las tareas de investigación del grupo de microgravedad del Laboratorio de Aerodinámica. Con su tesis doctoral, leída en 1981, se inició el estudio de la dinámica de puentes líquidos axilsimétricos, tema sobre el que ha publicado una decena de trabajos en revistas especializadas. Sanz entró a formar parte del equipo del Laboratorio de Aerodinámica en 1977. Experto en técnicas de experimentación con puentes líquidos en microgravedad simulada, desarrolló la teoría que soporta la validez de tales técnicas de experimentación en procesos dinámicos. Doctor ingeniero aeronáutico desde 1983, es el responsable del plan de ensayos en tierra, así como de los experimentos a realizar dentro del programa alemán TEXUS de cohetes de sondeo.

Hace...

José M.ª López Piñero

...cuatrocientos años

Juan Fragoso publicó la segunda edición de su *Chirurgia Universal*, uno de los principales tratados quirúrgicos aparecidos en la España del siglo XVI.

Nacido en Toledo hacia 1530, Fragoso estudió en la Universidad de Alcalá, donde se graduó de bachiller en medicina el año 1552. Ejerció en Sevilla hasta finales de 1570, fecha en la que se trasladó a Madrid como cirujano de la Casa Real. Allí permaneció el resto de su vida, primero al servicio de la reina Ana y, más tarde, al del propio Felipe II. Falleció en 1597.

Fragoso merece ser recordado como naturalista por sus estudios sobre la flora peninsular. Sabemos, por ejemplo, que acompañó a Francisco Hernández en varias excursiones botánicas por diversas zonas de Andalucía a mediados de los años cincuenta. Recogió los resultados de dichos estudios en su *Catalogus simplicium medicamentorum* (1566) y en la titulada *De succedaneis medicamentis* (1575), pero no llegó a terminar la *Hispanicarum plantarum Historia* que proyectaba. Publicó, en cambio, un *Discurso de las cosas aromáticas, árboles y frutales y de otras muchas medicinas simples que se traen de la India Oriental* (1572), basado principalmente en la obra del portugués García de Horta, al que no menciona, aunque sí a su traductor Carolus Clusius. También se ocupa de algunas plantas americanas estudiadas por Nicolás Monardes, al que tampoco cita. Algunas de ellas había podido verlas en el jardín botánico que Felipe II había fundado en Aranjuez y casi siempre recuerda su experiencia personal al administrarlas como medicamento. Este libro fue traducido al latín y publicado en Estrasburgo en dos ocasiones. Véase, como ejemplo, lo que dice acerca de la "raíz del Mechoacán" (*Ipamoea jalapa* Pursh): "Se descubrió en el Nuevo Mundo, tomando el apellido de una región del mismo nombre. Críase en Aquacatlán, pueblo de la Nueva Compostela o Galicia y no en Mechoacán, como algunos falsamente pensaron por razón del nombre, el cual tomó porque los que primero supieron la virtud, y lo usaron, fueron de la provincia de Mechoacán. Lo que se ha podido saber de su origen es que desta raíz salen unas hojas grandes y redondas, con una punta pequeña, y que echa unos racimos con unas uvillas del tamaño del calantro seco que maduran por el mes de Septiembre. Produce muchos ramos sobre la tierra que, si los ayudan, van trepando por lo alto a ma-

377

SVMA DE LAS PROPOSICIONES de Cirugia, que el Licenciado Fragoſo enſeña, contra vnos auiſos, que imprimó vn Doctor de eſta facultad el año de mil y quinientos y ochenta y quatro.

PRIMERA PROPOSICION.

DEzir, que en ninguna herida de cabeça ſe ha de legrar el caſco; aunque eſtè cortado, quebrado, ò contuſo, y aunque comprima, y hiera, ò punce el celebro, es vna ſecta muy
Differenti. 181. antigua, pero no racional, ſino impirica; que ſiguieron tambien algunos modernos, entre los quales fue Hugo, y Teodorico, y el Conciliador, vſando de poluos, y confecciones, y bebidas, que tuuieſſen propiedad de mundificar, reſolver, ù deſſecar, y conſumir qualquiera materia, no ſolamente en la cabeça, ſino en los miembros interiores del pecho, y vientre, y de ſanar fiſtulas, y llagas malignas. El ma- de temer, que en tan largo tiempo penetrarà la materia por èl haſta la duramater, lo qual ſe remedia cō horadar el caſco, para que ſalga lo que huuiere entrado. Y eſto es lo que dixo Galeno; que ſi ninguna coſa deſde las partes heridas cayeſſe dētro, ſeria ſuperfluo legrar el hueſſo. Es el tercero argumento, que ſi en las fracturas de los otros hueſſos, conuiene deſcubrir quando eſtà eſcondida alguna materia dentro dellos, quanto mas en el celebro; cuya corrupcion es la miſma muerte. Y aſsi Hypocrates, Principe de los Medicos racionales, conſiderando eſto, dixo las palabras ſiguientes, confirmadas con la experiencia.

1. Página inicial del primer capítulo del Discurso de las cosas aromáticas, árboles y frutales, y de otras muchas medicinas simples que se traen de la India Oriental *(1572), de Juan Fragoso*

nera de nuestros lúpulos. La simiente que nos enviaron estos años pasados es de dos maneras, la una negra y esquinada y muy dura, y la otra de la misma largura, más tierna, de color leonado, y encerrada en unos vasillos del mismo color, algo vellosos por dentro. Algunos pensaron, y no sin muy gran fundamento y autoridad, que esta raíz fuese nuestra nueza blanca vulgar, pero vemos que ésta, verde y seca, pica al gusto y el mechoacán no tiene ningún sabor. En los jardines de Su Majestad se ve nacido en alto, que va hilando sobre unas cañas con hojas de corregüela, o semejantes a las de la yedra. De cuya raíz hicimos ya experiencia este año, dando el polvo en la cantidad ordinaria a un criado nuestro, con que no solamente no purgó, pero tuvo muchas congojas y bascas, con otros accidentes malos; por donde parece haber ayudado poco al medicamento la disposición deste suelo. Después acá han traído de Tierra Firme otra suerte de Mechoacán cogido en la costa de Nicaragua y en Quito, que tiene la raíz menor con alguna agudeza, por el cual respecto no se tiene por tal como el primero. El mejor ha de ser el blanco, el no muy pesado y sin gusto. Consérvase entre mijo, o envuelto en un encerado. Es fácil de tomar por no tener mal gusto y, por tanto, es buena medicina para niños y viejos, y para los que no pueden tomar otra cosa. Purga principalmente flema y el agua de los hidrópicos. Conviene en la ictericia, en dolor de cabeça antiguo, para lamparones, para vaguidos, gota coral y corrimientos antiguos... Tómase el polvo medianamente molido en cantidad y peso de medio real, si es niño, si mozo, de uno, y al mayor, de dos. Háse de beber por la mañana desatado en vino blanco, y para el que no lo bebiere, en agua de hinojo, canela, o de anís, o en el vino aguado con agua de achicoria...".

Su labor de naturalista, sin embargo, no debe hacer olvidar que Fragoso fue, ante todo, un cirujano. Debido a su formación en Alcalá, junto a figuras como Francisco Valles, Cristóbal de Vega y Fernando de Mena, la base doctrinal de su obra quirúrgica fue el galenismo de orientación "hipocratista". Subrayó, por ello, la importancia de la observación clínica, tomando como modelo los textos hipocráticos. También tuvo gran interés por las nuevas corrientes del saber anatómico.

En 1570, publicó unos *Erotemas Chirurgicos*, que contenían "todo lo más necesario del arte de la cirugía". Once años más tarde, apareció la primera edición de su *Chirurgia Universal*, en la que reunió diversos escritos de tema anatómico, quirúrgico y terapéutico, algunos inéditos y otros reelaboraciones o simples reimpresiones de textos que había publicado anteriormente.

Incluye, en primer término, un compendio de cirugía dividido en los "libros" habituales de la época –anatomía, apostemas, heridas, úlceras, fracturas y dislocaciones– cada uno de ellos con una amplia "glosa". El contenido es notable, más que por aportaciones personales, por la rica y actualizada información que maneja Fragoso. Por ejemplo, fue el primer médico español que cita a Paracelso y, en su disposición anatómica, las obras de Vesalio quedaron ya claramente desplazadas por las de Realdo Colombo y Gabriele Falloppio. "Los murecillos (músculos) –dice en el capítulo dedicado a los ojos– son cinco; cuatro nacen dentro de la cuenca, junto a la raíz del nervio de la vista y acaban en medio del ojo, rodeándole todo, cada uno igualmente; son muy delgados, están los dos a la parte de arriba y los otros dos en la de abajo, cada uno de su lado; y aunque ayudan a sostener el ojo dentro del casco, sirven también de moverle arriba o abajo, o a los lados, cuando abren solos; pero cuando todos juntos, si tiran a la par, tenemos el ojo quedo, y derecho; si uno tras otro, movemos el ojo alrededor, sin que sea menester para este oficio otro particular murecillo, como Vesalio pensó. El uso del quinto, aunque fue para levantar el ojo hacia arriba, como cuando le echamos en blanco, también sirve de tenerle firme. Este, dice Realdo, que fue el primero que le halló, y llámale admirable, porque comienza del ojo y acaba en él, y que así es cosa dificultosa decir cuál sea su propio movimiento. Galeno y Vesalio dieron siete murecillos a cada ojo, pero los dos que añadieron sirven de mover los párpados sin tocar en él. Falopio dice hallarse otro más, cuyo oficio es mover y levantar hacia arriba el párpado más alto. Valverde da tres murecillos a los párpados; los dos primeros sirven de cerrar el párpado de arriba, el cual sólo menean los hombres, estando quedo el de abajo; el tercero le ayuda a abrir".

A partir de su segunda edición (1586), la obra incorporó una *Suma de proposiciones de Cirugia*, en la que dos años antes Fragoso había criticado la "vía particular", es decir, método curativo defendido por el sevillano Bartolomé Hidalgo de Agüero, acusándole principalmente de favorecer el abstencionismo quirúrgico; es especial en cirugía craneal: "Si fuese una fractura o contusión que actualmente no daña ni ofende al cerebro, podríase dejar el uso

DISCVRSO
PRIMERO DEL
AMBAR.


CERCA del origen y nacimiẽto d'l ambar, ay diuerſas opiniones. Vnos tienen por aueriguado ſer ſimiente de Valleña, o cierta opilacion criada en ſu buche, o alomenos eſpuma de la mar. Todo lo qual parece falſo, por no hallarſe donde ſe veẽ muchas vallenas, ni a do ſe allega gran ſuma de eſpumas, con el cõtinuo mouimiento de las olas. Aliende deſto, ſi fueſſe ſimiente de vallenas, todas ellas lo terniã

A · yes

2. *Cabecera de la* Suma de las proposiciones de Cirugía *(1586), en la que Fragoso criticó el método curativo de Bartolomé Hidalgo de Agüero*

de los instrumentos en confianza de las medicinas y pociones que dijimos; pero siendo el caso tan urgente como es cargar el casco subintrado y rompido sobre la tela de los sesos, dejar a beneficio de natura algún pedazo del mismo hueso que está picando, es grandísimo desatino y locura; porque no puede naturaleza reducir y levantar el casco hundido, si no fuese en los niños, por ser tiernos, y aún en éstos apenas y con ayuda de alguna ventosa o de algún emplastillo de los que escribimos, tratando de la contusión del casco. Ni tampoco puede naturaleza vomitar y sacudir de sí el hueso pungente; a lo cual se podría aguardar en otros miembros, pero en el cerebro ni en un solo punto, por ser tan principal y que no consiente inflamaciones, sino a grandísima costa y peligro".

La *Chirurgia Universal* contiene una traducción castellana comentada de los *Aforismos de Hypocrates, tocantes a la Cirugia*. Incluye también un *Tratado de las declaraciones que han de hazer los Cirujanos*, que es una de las primeras monografías sobre medicina legal, así como tres escritos de tema terapéutico.

Terapéutica y lucha contra el cáncer

Al juzgar el progreso en la guerra contra el cáncer debe comprenderse el papel desempeñado por el procedimiento experimental de ensayo y error en la evolución de la medicina y conocerse las formas más comunes de cáncer

John Cairns

El cáncer es una enfermedad que, directa o indirectamente, nos afectará a la mayoría. En Occidente, una de cada tres personas lo padece en algún momento de su vida, y una de cada cinco muere de ese mal. Mediremos en este artículo el progreso alcanzado en su tratamiento. Por ser materia controvertida, desentrañaremos los elementos que entran en la elaboración de los juicios que se emiten y abordaremos las fuentes de información disponibles. Antes de entrar en discusión, resultará de utilidad considerar brevemente el desarrollo de un tratamiento contra una enfermedad cualquiera.

En la biología molecular de todos los seres vivos se advierte una unidad subyacente, aun cuando la forma particular de cada una de las especies refleje acontecimientos fortuitos de su historia evolutiva. Tenemos la certeza de que cualquier nuevo animal que se descubra presentará las mismas macromoléculas portadoras de información y el mismo código genético que las otras formas de vida; sin embargo, aunque nos proporcionasen una descripción completa del hábitat del animal, no podríamos predecir exactamente su aspecto externo, que dependerá de la historia de sus antepasados. Y es todavía menor lo que podríamos predecir acerca de las enfermedades del animal. Por ejemplo, ¿quién podría imaginar que *Homo sapiens* comparte con el humilde cobaya la poco envidiable condición de ser incapaz de sintetizar ácido ascórbico, o que, al igual que los armadillos, es susceptible a la bacteria que causa la lepra, o que el cáncer intestinal suele afectar al hombre en el intestino grueso y a la oveja en el delgado?

Incapaces de predecir la existencia y las características más destacadas de cada enfermedad, carecemos de base que nos permita decidir cómo debe tratarse ésta. No parece que costase mucho, a partir de esa constatación, concluir que el tratamiento de toda enfermedad haya de apoyarse en algún sistema racional de ensayo y error. Sin embargo, esa noción es nueva en los anales de la medicina. A comienzos del siglo XIX se desarrolló un famoso estudio comparativo del destino de un grupo de pacientes con neumonía sometidos a sangrías a lo largo de diversos estadios de la enfermedad. Se comprobó que la pérdida de sangre no afectaba ni a la duración media de la enfermedad ni al desenlace fatal de la misma. La mayoría de los pacientes estuvieron enfermos durante dos o tres semanas y un 25 por ciento de ellos falleció. El autor del estudio no se aventuró a ir más lejos ni a sugerir que mejor hubiera sido dejar en paz a los enfermos, pero aun así se le atacó por haberse atrevido a pensar que cabía comparar un paciente con otro. Como afirmaba uno de sus críticos: “Al invocar la inflexibilidad de la aritmética para huir de las intrusiones de la imaginación, se comete un ultraje al buen sentido”. (En ausencia de pruebas clínicas adecuadas, el informe se consideró dudoso y, hasta bien entrado el siglo XX, no perdió favor la sangría.)

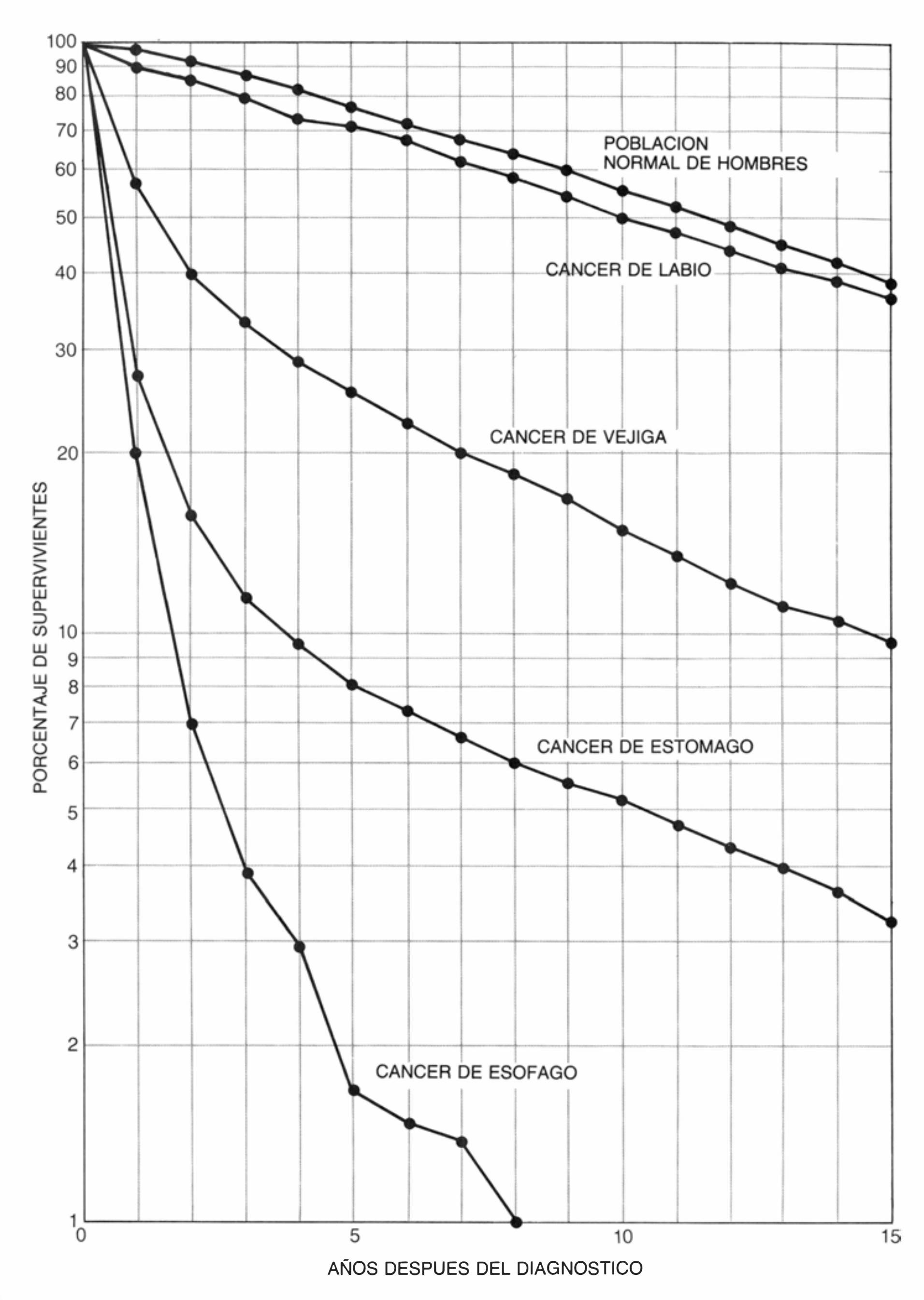
En las enfermedades del tipo de la neumonía o el cáncer, unas veces fatales y otras no, no parece haber forma de determinar si un determinado paciente estaba predestinado a sobrevivir o si su supervivencia debióse al tratamiento. Ello obliga a comparar la respuesta de un grupo de pacientes, y no limitarse a la respuesta de uno o dos individuos. Se ha reconocido la existencia de más de 100 clases diferentes de cáncer. Cada una de ellas posee un desarrollo característico, definido por la edad media de los sujetos en los que aparece, el ritmo de crecimiento y la tendencia a extenderse y a ser letal. Por ello, cada clase debe considerarse una enfermedad distinta. Además, la mayoría de los pacientes de cáncer son de mediana edad, o ancianos, lo que obliga a distinguir ésta de otras causas de muerte “competidoras”. Después de todo, difícilmente podría esperarse que ni siquiera los mejores tratamientos protejan a un paciente de 90 años frente a todas las formas de mortalidad. Como en el caso del médico del siglo XIX que estudió la neumonía, el clínico moderno comienza a investigar qué les sucede a los pacientes de cada tipo de cáncer y cuál es su esperanza de vida en comparación con la de personas libres del mal.

Aparece el cáncer cuando una célula del cuerpo empieza a multiplicarse sin restricción y produce una familia de descendientes que invaden los tejidos circundantes. La invasión puede ir seguida de metástasis, o dispersión a puntos distantes, a través del sistema linfático y el torrente circulatorio. El proceso de metástasis es la razón principal de la mortalidad del cáncer, porque lleva la enfermedad más allá del alcance de la cirugía y de la irradiación locales. Algunos cánceres, por razones desconocidas, son incapaces de producir metástasis (por ejemplo, la mayoría de las formas de cáncer de piel). Se manejan mejor estos tipos, a menos, claro está, de que el proceso de invasión local sea por sí mismo letal (como ocurre en ciertas formas de cáncer de cerebro). En el otro sistema, las células normales de la médula ósea y del sistema linfático están ya progra-

madas para recorrer el organismo a todo lo largo y ancho, por lo que no sorprende que los cánceres de esas células (las leucemias y los linfomas) se diseminen desde su misma iniciación. La mayoría de las formas de cáncer se encuentran entre esos dos ejemplos extremos.

Desde hace poco, podemos estudiar directamente los cambios de la estructura y función de los genes de ciertas células cancerosas [véase "Base molecular del cáncer", por Robert A. Weinberg; INVESTIGACIÓN Y CIENCIA, enero de 1984]; sin embargo, apenas sabemos algo acerca de qué es lo que controla la multiplicación y la restricción territorial de la mayoría de las células y tejidos. En estos momentos, cuanto conocemos del comportamieno y pronóstico de cada tipo de cáncer sigue teniendo, en su mayor parte, carácter empírico. A lo largo de lo que va de siglo, los patólogos han elaborado una clasificación de los cánceres humanos, de acuerdo con el origen y la categoría de las células implicadas, y posteriormente han subdividido cada tipo de acuerdo con el aspecto de las células y su pauta general de crecimiento. El valor de dicha clasificación reside en que los diferentes cánceres se comportan de formas muy distintas. Algunos provocan la muerte en poco tiempo y otros no; la mayoría se presenta predominantemente en la edad madura, pero los hay que sólo se dan en los niños. Comunes unos; raros, otros. Algunos son más frecuentes en las naciones ricas y menos en las pobres, y en otros ocurre lo contrario.

A partir de la segunda guerra mundial, y con el fin de recoger y almacenar las cambiantes tendencias de la incidencia y mortalidad del cáncer, se han establecido registros en varios estados y naciones. Los datos recogidos por algunos de esos registros ofrecen una imagen bastante precisa de la historia natural del cáncer y ése habrá de ser el necesario punto de partida para cualquier tipo de discusión sobre el tratamiento. El Registro del Cáncer de Noruega ha publicado sus hallazgos en forma de libro, lo que lo convierte en fuente de datos muy accesible. Noruega tiene una población aproximada de 3,5 millones de habitantes; el registro sigue el destino de 220.000 casos de cáncer diagnosticados a partir de 1953. Como muestra de tales estadísticas, la figura 1 presenta la pauta de supervivencia de hombres con cuatro clases de cáncer, elegidos para demostrar la diversidad de comportamientos de estas



1. SUPERVIVENCIA DE VARONES NORUEGOS aquejados de diversos tipos de cáncer, comparada con la supervivencia de sujetos masculinos de la misma edad del conjunto de la población nacional. En los años siguientes al diagnóstico, el cáncer de labio produjo una despreciable disminución de la esperanza de vida, mientras que el cáncer de esófago solía conducir rápidamente a un desenlace fatal. Las tasas de supervivencia registradas en pacientes con cánceres de vejiga y de estómago se situaban entre esos dos extremos. Las estadísticas cubren los años que van desde 1953 hasta 1964 y proceden del Registro del Cáncer de Noruega.

afecciones. En algunas localizaciones, como en el labio, el cáncer va asociado a un imperceptible descenso de la esperanza de vida; en otras, como el esófago, el cáncer casi siempre resulta fatal en breve tiempo.

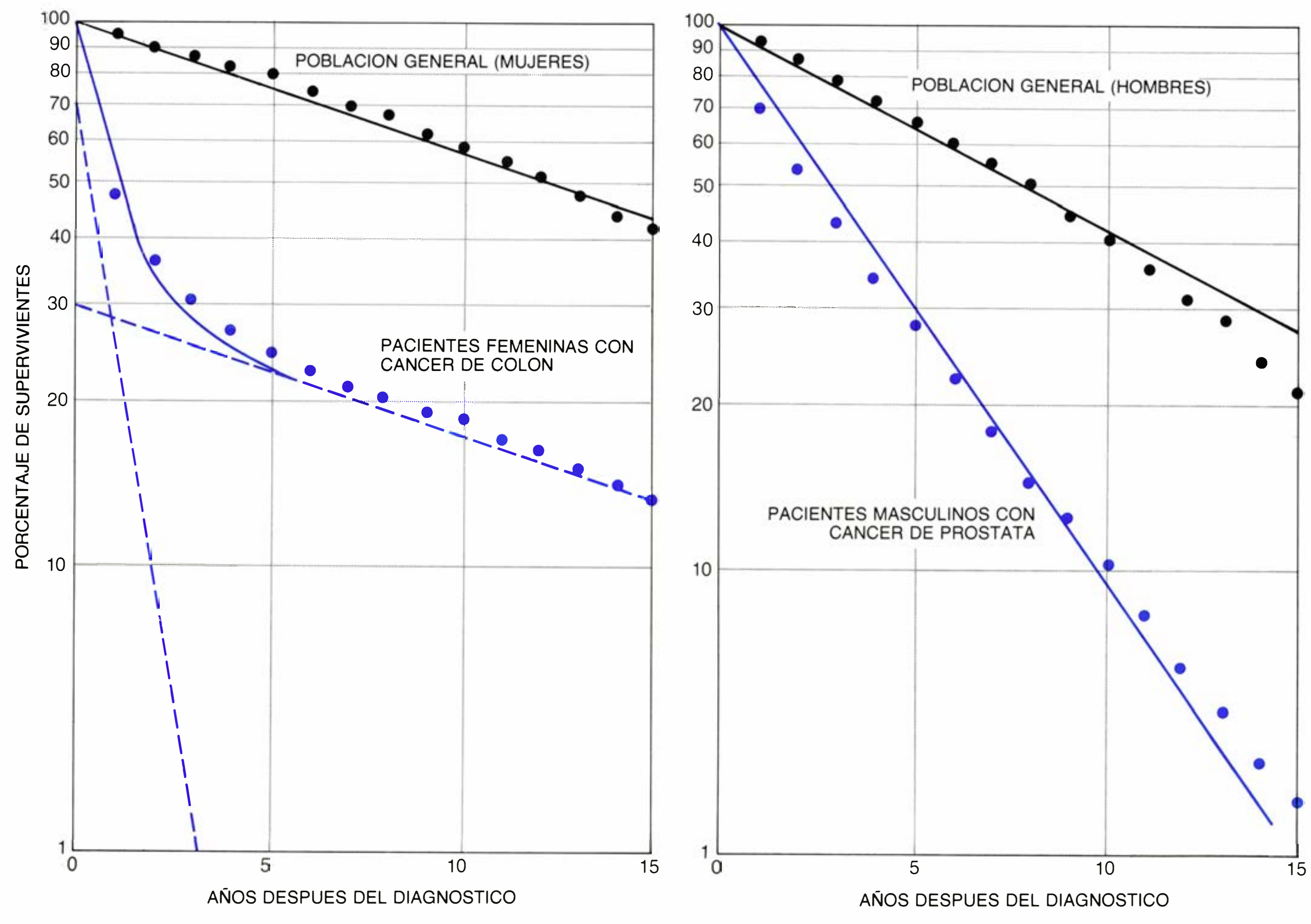
Un grupo de pacientes de cáncer puede considerarse curado si sus integrantes fallecen con una tasa de mortalidad igual que la población general, esto es, si gracias al tratamiento han retornado al grupo general. Por tanto, el paso esencial al determinar la frecuencia con que puede curarse o controlarse un tipo concreto de cáncer es calcular, fundándonos en las estadísticas, qué proporción de pacientes muere a la misma edad que les correspondería de no haber sufrido cáncer (es decir qué fracción muere por causas no relacionadas con su cáncer). Por ejemplo, se ha comparado la tasa de supervivencia de las mujeres noruegas con cáncer de colon con la tasa de supervivencia de la población general de mujeres caracterizada por la misma distribución de

edades. La mayoría de las pacientes murieron poco tiempo después del diagnóstico, pero una minoría apreciable, alrededor del 30 por ciento, se ajustó a la tasa de la población general (es decir, se comportó como si se hubiese curado). Eso es lo que cabría esperar si algunas de las pacientes murieran del cáncer y otras no.

El registro noruego cita 40 localizaciones principales de cáncer en los varones y 43 en las mujeres. Cálculos similares a los anteriores pueden hacerse para cada una de las localizaciones. Al sumar los datos estimados, se infiere que aproximadamente el 25 por ciento del total de pacientes varones con cáncer y alrededor del 40 por ciento de las mujeres con cáncer murieron por motivos distintos del cáncer. Dicho de otra forma, aproximadamente una tercera parte del total de pacientes noruegos con cáncer no vio acortada su vida como resultado de su enfermedad.

En estas estadísticas de las décadas de 1950 y 1960 se recogen los resultados del tratamiento con cirugía, ocasionalmente respaldado por irradiación con rayos X cuando el tumor primario resultaba innacesible. Es la imagen de lo que resultaba habitual antes de la llegada de los programas de rastreo, o *screening*, de la quimioterapia y de numerosas pruebas clínicas. Las principales ayudas complementarias a la cirugía, tales como las transfusiones sanguíneas, los antibióticos y las formas mejoradas de anestesia, ya se habían desarrollado y propagado a principios de la década de los 50, por lo que el dato decisivo para casi todos los pacientes era, por esas fechas, la cuantía en que se había extendido el cáncer en el momento de la intervención quirúrgica. Si el cáncer había producido metástasis en localizaciones inaccesibles a la cirugía o a la radioterapia, el destino del paciente estaba determinado por la tasa de crecimiento y posterior dispersión de las restantes células cancerosas. Sólo si intervenía alguna otra causa de muerte se libraría el paciente de morir de cáncer.

La importancia de la cuantía de la dispersión del cáncer en el momento de la operación se recoge en la figura 3, que presenta la superviviencia de mujeres con cáncer de colon, de acuerdo con el estado de su enfermedad. Sencillamente, las oportunidades de vivir de una paciente eran peores si su cáncer ya se había extendido en el momento del primer diagnóstico. Existen, sin embargo, dos posibles explicaciones de ese efecto. La primera y más obvia: el momento en que se emite el diagnóstico resulta ser determinante. Según



2. SE CONSIDERAN CURADOS LOS PACIENTES DE CANCER si mueren con la misma tasa de mortalidad que la población general. La gráfica izquierda compara la tasa de supervivencia de mujeres noruegas afectadas de cáncer de colon (*puntos de color*) con la de la población general de mujeres con la misma distribución de edades (*puntos negros*). La línea continua gris muestra que la población general tenía una vida media de 13 años. (La vida media es el tiempo requerido para que muera la mitad de la población.) La línea continua de color es la suma de las dos líneas de puntos, y muestra que cabe encuadrar las mujeres con cáncer de colon en dos categorías: el 70 por ciento presentaba una vida media de ocho meses y, el 30 por ciento, de 13 años; un subgrupo equivalente al 30 por ciento de las mujeres moría con la misma tasa de mortalidad que la población general. Ciertos cánceres, por el contrario, no presentan subgrupos, y en este sentido podrían considerarse incurables por los métodos actuales. El cáncer de próstata constituye un ejemplo; la gráfica derecha compara la tasa de supervivencia de hombres noruegos afectados de cáncer de próstata (*puntos de color*) con la tasa de supervivencia de hombres de la misma edad de la población general (*puntos negros*). La línea negra muestra que la población general tenía una vida media de ocho años y, la línea de color, que los pacientes con cáncer mostraban una vida media de tres años. Alrededor de un tercio de los pacientes con cáncer no sufrió acortamiento de su vida a consecuencia de su enfermedad.

ese punto de vista, los cánceres que han producido ya metástasis cuando se descubren permanecieron abandonados demasiado tiempo y debieron haberse diagnosticado antes, cuando aún estaban localizados.

La explicación menos obvia es que lo que aquí se refleja no es una variación del tiempo de diagnóstico, sino una variación de la capacidad de extenderse el cáncer y producir metástasis. Con otras palabras, el cáncer localizado podía estar destinado a permanecer localizado durante muchos años después de hacerse detectable, en tanto que el cáncer que había producido metástasis quizá constara de células tan aptas para su dispersión que ya había producido metástasis cuando se hallaba en ciernes y escapaba a toda detección. Si la primera explicación es la correcta, el diagnóstico precoz resultaría muy beneficioso; si la segunda, no existiría beneficio apreciable. Las ventajas reales de un diagnóstico precoz deben determinarse, por tanto, mediante ensayos adecuadamente controlados, y ello para cada tipo de cáncer, pues lo que es cierto para el cáncer de mama quizá no lo sea para el de pulmón.

Un famoso ejemplo de pruebas encaminadas a medir el efecto del diagnóstico precoz la tenemos en el estudio del cáncer de mama iniciado hace 20 años en Nueva York, subvencionado por el Instituto Nacional del Cáncer. Se acometió el seguimiento de 62.000 mujeres, de edades comprendidas entre los 40 y los 60 años, acogidas al Plan de Asegurados de la Salud del Gran Nueva York. Se encuadraron las mujeres en varias categorías, de acuerdo con su edad, tamaño de su familia y renta. Cada categoría se dividía luego al azar en dos grupos. A las de un grupo, llamado "de ensayo" (formado por 31.000 mujeres), se les ofrecía una revisión anual gratuita, consistente en un examen físico y una mamografía con rayos X, para la detección precoz de cáncer de mama. A las 31.000 restantes, el grupo "control", no se les ofreció ninguna ayuda especial. Las revisiones anuales del grupo de prueba se prolongaron durante cuatro años y, desde entonces, se lleva un control monitorizado del posterior destino de ambos grupos.

La prueba experimental se diseñó, por tanto, con el fin de dar respuesta a la siguiente cuestión práctica: ¿rinden algún beneficio apreciable revisiones periódicas gratuitas? (Es decir, ¿pueden atajarse los cánceres de mama antes de que se hayan dispersado, y será suficientemente elevada la proporción de mujeres que consientan el examen?) Las respuestas fueron prometedoras. Dos terceras partes de las mujeres del grupo "de ensayo" se habían sometido al menos a un examen y, durante los nueve años de seguimiento, el grupo de ensayo, en conjunto, sufrió un número significativamente menor de muertes por cáncer de mama (91 frente a las 128 del grupo de control).

El estudio demostró, además, la importancia de la selección aleatoria de los grupos de control. Por encima de todo, los dos grupos mostraron aproximadamente la misma incidencia de cáncer de mama y la misma mortalidad general debida a causas distintas del cáncer, tal como debía suceder, ya que los grupos se seleccionaron al azar. Hubo un hallazgo notable: en el grupo de ensayo, las 10.200 mujeres que rehusaron el examen presentaron una mortalidad general más elevada, aunque una incidencia ligeramente inferior de cáncer de mama y de muerte por esa causa, que las 20.000 mujeres que aceptaron ser examinadas.

La explicación radica en el hecho de que la separación del grupo de prueba entre examinadas y no examinadas fue el resultado de una autoselección. Las mujeres que rehusaron el control estaban menos interesadas por su salud y demostraron un nivel de educación inferior al de las mujeres que lo aceptaron. Dado que el cáncer de mama se da más entre las mujeres cultas y bien situadas que entre las indigentes y de menor cultura, esas diferencias entre las dos categorías autoseleccionadas no resultan sorprendentes. De donde se desprende una gran lección. Si la comparación se hubiese establecido simplemente entre mujeres examinadas y las que rehusaron a serlo, el estudio podía haber conducido a la ridícula conclusión de que el examen anual para detectar los signos precoces del cáncer de mama reduce la mortalidad general y eleva ligeramente la mortalidad debida a la propia enfermedad que se trataba de prevenir mediante ese ejercicio. Por incluirse un control seleccionado al azar, el estudio produjo una estimación fiable de los beneficios reales del examen en busca de posible cáncer de mama.

En resumen, los resultados del estudio realizado por el Plan de Asegurados de la Salud de Nueva York (y los de un ensayo sueco semejante recientemente publicado) sugieren que una cuarta parte aproximada de la mortalidad total por cáncer de mama (es decir, la cuarta parte de unas 35.000 muertes anuales) podría prevenirse si a todas las mujeres norteamericanas de edad superior a los 50 años se les ofreciese un examen gratuito cada uno o tres años. El presupuesto del programa (más de 100 millones de dólares al año) ha impedido su realización, pero al menos quedan ahí los datos, dispuestos para su empleo en cualquier cálculo o estimación de prioridades, costos y beneficios.

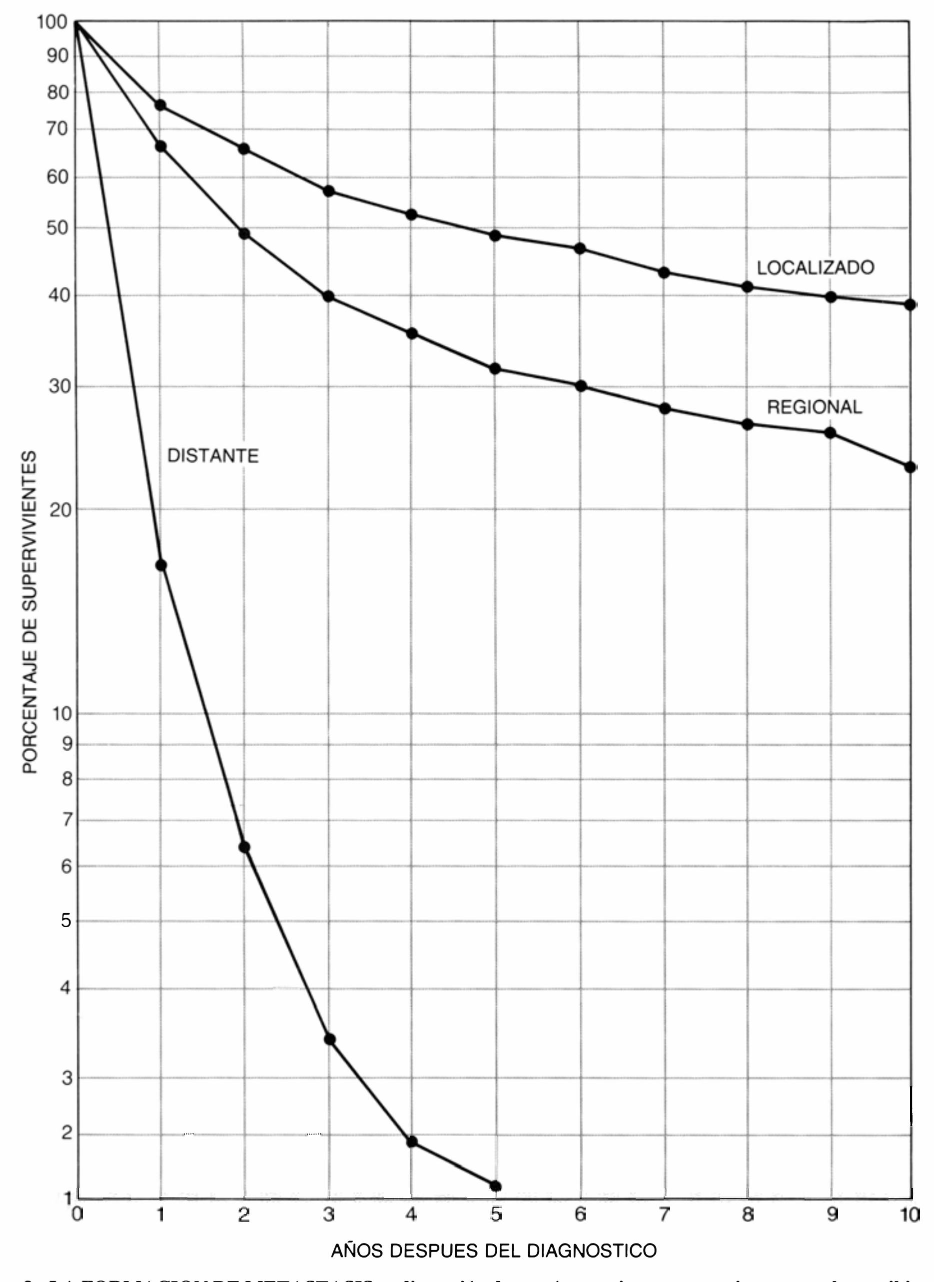
El otro procedimiento principal de rastreo es el frotis de Pap, o prueba de Papanicolaou, para el diagnóstico precoz de cambios precancerosos del cérvix. En términos generales, y en todo el mundo, el carcinoma de cérvix es el cáncer letal más común entre las mujeres. Como muchas otras formas de cáncer, pero de manera distinta del de mama, predomina entre las mujeres pobres y de baja cultura. En Estados Unidos, por ejemplo, las tasas entre las clases sociales más altas y más bajas difieren en un factor de cinco.

El frotis de Pap lo inventó, a finales de la década de 1920, Aurel Babès, de Bucarest, y lo desarrolló George N. Papanicolaou, en la Facultad de Medicina de la Universidad de Cornell. La prueba consiste en el examen microscópico de células raspadas de la suerficie del cérvix, en el punto de entrada al útero. Durante los años siguientes a la segunda guerra mundial, empezó a utilizarse para detectar los estadios precoces del desarrollo del cáncer de cérvix. En la actualidad se ha practicado la prueba una o más veces al menos al 75 por ciento de las mujeres occidentales adultas. No se ha intentado un ensayo adecuadamente controlado, como el que se hizo para el cáncer de mama, razón por la cual no se han podido establecer los beneficios de la prueba. Ciertamente la empresa resulta complicada.

La mortalidad provocada por cáncer de cérvix en Estados Unidos ha ido disminuyendo gradualmente, tal vez desde los años 1930, por la probable razón de que, durante ese período, han aumentado los niveles higiénicos, económicos y educativos. Dado que el frotis de Pap se introdujo en Estados Unidos cuando el cáncer de cérvix ya disminuía, no pueden utilizarse las cifras de mortalidad total nacional como prueba de su éxito. Ni tampoco cabe comparar la mortalidad entre las mujeres que se habían sometido a la prueba con la de aquellas que no lo ha-

bían sido; en ausencia de una fuerte presión, las más educadas accederán con más facilidad a la prueba que las de menor cultura, por lo que cabe esperar que una mortalidad menor por cáncer de cérvix aun cuando la prueba no ofrezca ningún beneficio.

La cuestión de la eficacia del frotis de Pap se basa en dos observaciones independientes. En primer lugar, siempre que pueden compararse poblaciones de mujeres, en todo lo demás iguales, a las que se ofreció la posibilidad de someterse a programas de pruebas desarrollados en distintos momentos (por ejemplo, mujeres de diferentes provincias de Canadá o de cada uno de los Países Escandinavos), parece que el descenso de la mortalidad por cáncer de cérvix se acelera coincidiendo invariablemente con la extensión de la práctica del ensayo. En segundo lugar, las mujeres que se encuentran en los estadios tempranos del cáncer de cérvix, pero que no se someten a tratamiento, acusan una mortalidad mucho más elevada por cáncer de cérvix que las que sí se someten a él. Ninguno de esos argumentos es por sí solo indiscutible, pero juntos adquieren solidez suficiente para que no resulte ético someter a un juicio adecuado el frotis de Papanicolaou.



3. LA FORMACION DE METASTASIS, o dispersión de un cáncer, ejerce un gran impacto en las posibilidades de supervivencia a largo plazo después del diagnóstico. Los datos que se muestran aquí proceden de un estudio realizado sobre mujeres noruegas de edades comprendidas entre los 55 y 74 años a las que se había diagnosticado un cáncer de colon. En algunas pacientes (*curva superior*) el mal estaba localizado en la pared intestinal; dos tercios de esas mujeres no parecen haber sufrido pérdida de su esperanza de vida. Cuando el cáncer se había dispersado por las glándulas linfáticas locales (*curva central*), sólo una tercera parte de las pacientes mantuvo una esperanza de vida normal. Si el cáncer había producido metástasis en localizaciones distantes (*curva inferior*) las perspectivas de supervivencia a largo plazo eran sumamente desfavorables.

En ocasiones, se ha intentado establecer programas para el diagnóstico precoz de otras clases de cáncer. Han demostrado su eficacia los de rastreo de cáncer de piel y de boca. Desgraciadamente, pocas de las localizaciones de los cánceres letales más comunes resultan tan accesibles como la mama, el cérvix, la piel y la boca, y, de momento, no hay indicios de que valga la pena aplicar a gran escala otros programas. En principio, los programas de exploración sólo pueden tener éxito en el caso de los cánceres caracterizados por un estado precanceroso precoz y detectable o los que suelen atravesar un estadio prolongado durante el cual se manifiestan sin que se hayan dispersado aún a zonas fuera del alcance de la cirugía.

Lamentablemente, no todos los tipos de cáncer se encuadran en alguna de esas categorías. Hace unos años, por ejemplo, un ensayo a gran escala para el diagnóstico precoz del cáncer de pulmón apuntó que no reporta beneficio alguno la detección de la enfermedad por radioscopia de tórax antes de la observación de síntomas; parece que cuando resultan detectables, la mayoría de los cánceres de pulmón se han dispersado ya demasiado para llegar a tiempo con el tratamiento. Por tanto, en cada tipo de programa de rastreo importa desarrollar antes un ensayo adecuado sobre un grupo de sujetos seleccionados al azar, puesto que no existe vía intuitiva que informe de cuáles son los tipos de cáncer susceptibles de interceptación mediante diagnóstico precoz.

En resumen, los programas de exploración en pos de un diagnóstico precoz aportan unas veces ventajas, pero otras no. Eficacia aparte, parece poco probable que ni siquiera los países más ricos de Occidente afronten los gastos que supone someter a pruebas anuales a la mayor parte de su población para detectar los signos precoces de todos los principales tipos de cáncer.

Sigue siendo triste realidad que menos del 50 por ciento de los pacientes pueden curarse gracias a la cirugía. Se desarrolla, por tanto, un tremendo esfuerzo en busca de formas coadyuvantes de tratamiento aplicables tras el

			MUERTES DEBIDAS A OTRAS CAUSAS EN LOS CINCO PRIMEROS AÑOS		CANCERES DE MAMA DETECTADOS EN LOS CINCO PRIMEROS AÑOS		MUERTES POR CANCER DE MAMA EN LOS NUEVE PRIMEROS AÑOS	
		NUMERO DE MUJERES DE CADA GRUPO	NUMERO	MUERTES POR CADA 1000 MUJERES	NUMERO	INCIDENCIA POR CADA 1000 MUJERES	NUMERO	MUERTES POR CADA 1000 MUJERES
GRUPO DE PRUEBA	EXAMINADAS	20,800	421	20	225	11	63	3.0
	REHUSARON	10,200	429	42	74	7	28	2.8
	TOTAL	31,000	850	27	299	10	91	2.9
GRUPO CONTROL	TOTAL	31,000	877	28	285	9	128	4.1

4. RASTREO PARA EL DIAGNOSTICO PRECOZ de cáncer de mama. Según un estudio financiado por el Instituto Nacional del Cáncer de los Estados Unidos, el rastreo (o *screening*) preventivo intercepta a veces la enfermedad cuando aún no se ha dispersado. En este estudio se hizo el seguimiento de 62.000 mujeres, de edades comprendidas entre los 40 y 64 años, acogidas al Plan de Asegurados de la Salud del Gran Nueva York. Las mujeres se separaron al azar en dos grupos: uno de ensayo y otro de control. A las mujeres que componían el grupo de prueba se les ofreció una revisión anual gratuita mediante examen físico y mamografía de rayos X para la detección precoz del cáncer de mama. Alrededor de dos tercios de esas mujeres aceptaron someterse al menos a un examen a lo largo de un período de cuatro años. A las del grupo de control no se las animó a someterse a ningún tipo de examen. Al cabo de nueve años de seguimiento, el grupo de prueba sufrió significativamente menos muertes debidas a cáncer de mama (91 muertes) que el de control (128 muertes).

acto quirúrgico. Las tres formas de terapia complementaria más utilizadas son el tratamiento con hormonas, la irradiación con rayos X y la quimioterapia.

El tratamiento hormonal es la forma obvia de terapia a aplicar en los cánceres de tejidos que responden a hormonas, como la mama y la próstata. En la década de 1890 empezó a extirparse los ovarios de mujeres que sufrían cáncer de mama en dispersión, por si la consiguiente caída de estrógenos circulantes retrasase el crecimiento de las células cancerosas. Para lograr el mismo efecto, sin eliminar los ovarios, se emplean ahora ciertos análogos estructurales de estrógenos, tales como el tamoxifén, que bloquean los receptores del estrógeno de las células cancerosas. Del mismo modo, a menudo se retrasa o inhibe el cáncer de próstata amputando los testículos o administrando estrógenos. Aunque no todos los cánceres de mama y de próstata responden al control hormonal, una ventaja de esta forma de terapia auxiliar es que sus efectos secundarios suelen ser leves.

Poco después de que Wilhelm Roentgen descubriese los rayos X, en 1895, se advirtió que la radiación podía dañar los tejidos humanos. En consecuencia, pronto empezó a aplicarse los rayos X en el tratamiento de los cánceres de mama que mostraban recurrencia local tras su extirpación quirúrgica. La irradiación con rayos X constituye hoy uno de los principales recursos de la terapia del cáncer. Pero hay que andar con cuidado. Una excesiva irradiación del cuerpo entero daña el sistema inmunitario, la médula ósea y el recubrimiento interno del intestino. El daño que sufren esos tejidos es la razón de los trastornos que provoca la radiación. (Actualmente interpretamos la mayoría de los efectos de la irradiación a través de la lesión del material genético, el ADN, y posiblemente ese daño tienda a ser mayor en los tejidos cuyas células se dividen rápidamente, porque cuanto mayor es la frecuencia con que se divide una célula, menor es el tiempo de que dispone para reparar cualquier lesión producida en su ADN.) Así, el tratamiento del cáncer por aplicación de rayos X depende de la sensibilidad relativa del tumor con respecto a la de los tejidos normales que lo rodean, así como de la posibilidad de concentrar la radiación sobre la zona maligna.

Conocemos hoy con toda precisión la cantidad de radiación que tolera cada parte del cuerpo. Además, disponiéndose ya de rayos X de alta energía, puede concentrarse con bastante exactitud la radiación en cualquier órgano elegido. Ello faculta el tratamiento de los cánceres más sensibles, como la enfermedad de Hodgkin, el cáncer de cérvix y una clase de cáncer testicular, sin producir un nivel inaceptable de trastornos debidos a la radiación. La mayoría de los cánceres, sin embargo, no pueden curarse por irradiación, pues la dosis de rayos X requerida para matar todas las células cancerosas acabaría también con el paciente.

El tratamiento que sigue en importancia a los rayos X, la quimioterapia citotóxica, parte de la observación de que uno de los efectos tóxicos a largo plazo de los gases mostaza utilizados en la primera guerra mundial era la lesión de la médula ósea. (Por cierto, exactamente igual que los rayos X, estos productos químicos tóxicos resultaron ser muy reactivos y capaces de dañar el ADN.) Poco después de la segunda guerra mundial, cuando el poder mutagénico de los gases mostaza dejó de considerarse un secreto militar, se emprendió una serie de pruebas encaminadas a ensayar la eficacia de esos productos químicos radiomiméticos (sustancias químicas que producen efectos similares a los de la radiación) en el tratamiento del cáncer. Los resultados iniciales fueron alentadores y, durante los 20 años siguientes, se añadieron muchos más productos químicos a la lista de drogas utilizadas en quimioterapia.

Hoy se emplea gran número de agentes quimioterapéuticos, en distintas combinaciones, para el tratamiento del cáncer. Algunos se preparan por síntesis química (recordemos la ciclofosfamida, ciertas nitrosoureas y, más recientemente, algunos compuestos orgánicos de metales, como el *cis*-platino). Otros son toxinas naturales (por ejemplo, alcaloides de plantas, como la vincristina, y toxinas de hongos, como

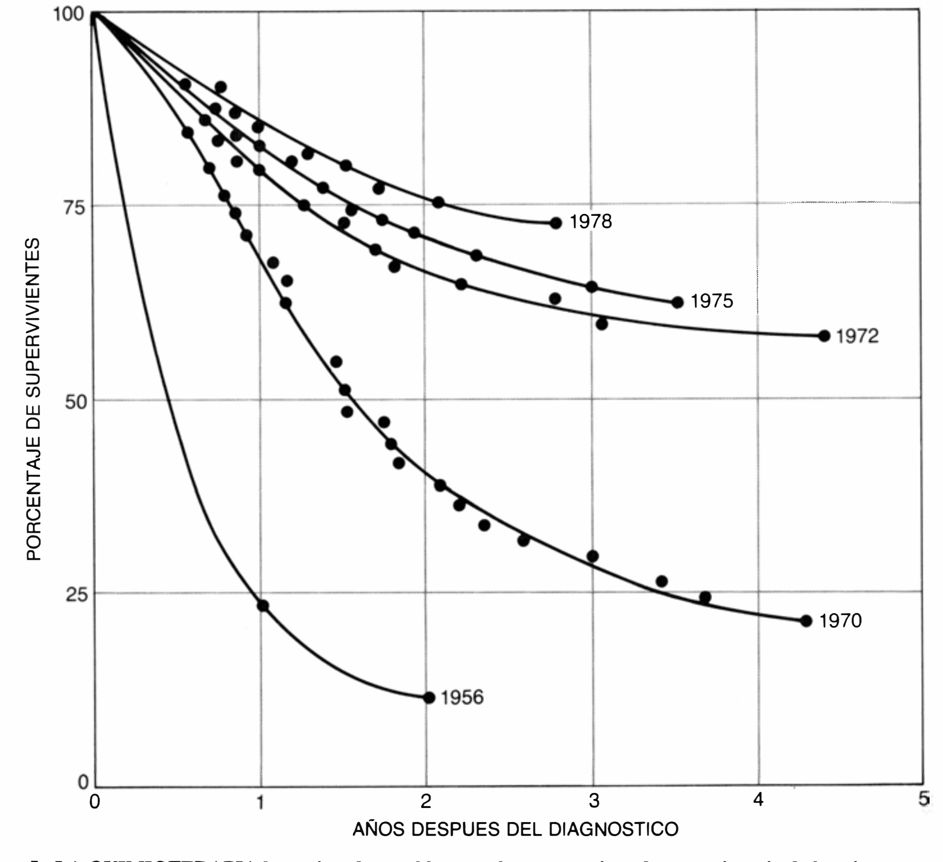
las actinomicinas). Casi todos se unen al ADN y provocan un daño que la célula no puede reparar adecuadamente. Ese parece ser el fundamento de su toxicidad, tanto para las células cancerosas como para los tejidos normales del cuerpo. El otro grupo de productos químicos habitualmente utilizados son ciertos antimetabolitos que bloquean la síntesis de ADN o de sus precursores (así, el fluoruracilo, el arabinósido de citosina y el metotrexato).

Los primeros esfuerzos en el campo de la quimioterapia se concentraron en leucemias y linfomas infantiles. Por dos razones. En primer lugar, porque esos cánceres, dispersos desde su inicio, resultaban inevitablemente fatales. En segundo lugar, porque los pacientes, jóvenes, tenían mucho más que ganar con su curación que las personas mayores. Aunque resultaba bastante fácil lograr una remisión temporal con quimioterapia, en casi todos los casos recurría el cáncer, mostrándose entonces más resistente a ulteriores quimioterapias.

En realidad, no debiera haber sorpresa en ello cuando hasta el menor cáncer susceptible de detección consta de al menos mil millones de células y se presume que cualquier población de células de ese tamaño contenga algunas variantes que crecen mejor frente a una presión selectiva. Partiendo de esas premisas se han establecido dos principios que, acertada o erróneamente, han determinado el desarrollo posterior de la mayoría de las clases de quimioterapia. Cuando de destruir todas las células de un cáncer se trata, quizá no sólo sea necesario el uso de los agentes químicos al máximo nivel tolerable, sino también la aplicación simultánea de agentes diversos.

Con combinaciones adecuadas de productos químicos se curan hoy muchos tipos de cánceres infantiles que, de otro modo, conducirían a un irremisible desenlace fatal. Por ejemplo, la mayoría de los niños con leucemia se curan, al menos aparentemente; para ser más precisos, una minoría recae y muere durante el tratamiento quimioterapéutico o poco después de finalizado éste, pero la mayoría entra en lo que, cuando menos, constituye un período prolongado de supervivencia sin recaídas. Cabe esperar que su esperanza de vida sea la normal. Resultados similares se han obtenido con otros cánceres de la infancia. Las estadísticas nacionales de mortalidad constituyen el mejor indicador del éxito. A comienzos de la década de 1950, en Estados Unidos morían cada año de cáncer unos 1900 niños menores de cinco años. Hoy son menos de 700 niños por año, lo cual sugiere que dos terceras partes del total de niños que han contraído el mal se curan.

La reducción de la mortalidad anual en niños mayores y en adultos jóvenes ha sido menos espectacular, con notables excepciones. Así, de la enfermedad de Hodgkin, que solía conducir inevitablemente a la muerte, se cura ya la mayoría de los pacientes. Unas mil vidas anuales, considerando todos los grupos de edad, se salvan hoy en Estados Unidos. Alrededor del 35 por ciento de los cánceres testiculares resultaban fatales antes del uso de la quimioterapia, en tanto que ahora puede evitarse aproximadamente una tercera parte de estas muertes, esto es, unas 300 vidas cada año. Finalmente, el coriocarcinoma, un cáncer raro de la placenta, que aparece aproximadamente en uno de cada 40.000 embarazos en Occidente, sana hoy administrando agentes quimioterapéuticos, con lo que se salvan de 20 a 30 vidas cada año. En conjunto, sin embargo, los logros son limitados. Las últimas cifras relativas a Estados Unidos arrojan unas 7000 muertes anuales por cáncer en pacientes de menos de 30 años, frente a los 10.000 que cabría esperar si la tasa de mortalidad no hubiese cambiado desde los años 1950.



5. LA QUIMIOTERAPIA ha mejorado notablemente las perspectivas de supervivencia de los niños que sufren leucemia. Sólo el 10 por ciento de los niños a los que se les había reconocido en 1956 la enfermedad seguían vivos dos años después; en 1978, el número de supervivientes a dos años era de aproximadamente el 70 por ciento. Los datos proceden de Dennan Hammond, del Grupo de Estudio de Oncología Infantil.

Hemos de señalar que, hasta ahora, no existen discrepancias entre el número total de curaciones calculado a partir de las diversas tasas de curación para ciertos cánceres y el cambio real recogido en el inventario nacional de muertes registradas. Cada año se curan, gracias a la quimioterapia, 3000 pacientes de edad inferior a 30 años; sin ella hubieran fallecido.

Sin embargo, sólo el dos por ciento de los pacientes que mueren de cáncer tienen menos de 30 años. En la gran mayoría de los cánceres, que se dan entre pacientes de más edad, los resultados de la quimioterapia son mucho más controvertidos. Las cifras de mortalidad reunidas y publicadas por el Instituto Nacional del Cáncer de los Estados Unidos muestran varios cambios importantes acaecidos durante los últimos 25 años. Las muertes por cáncer de pulmón van en aumento, particular-

mente entre las mujeres, como resultado tardío del incremento del hábito de fumar. Las muertes por cáncer de cérvix están disminuyendo, en parte, gracias a la prueba del frotis de Papanicolaou. La tasa de muerte por cáncer de estómago continúa su inexplicable tendencia descendente, que se inició en los años 1930; otros tipos, menos comunes, se inclinan ligeramente hacia una u otra dirección. Aparte del éxito logrado con la enfermedad de Hodgkin, la leucemia infantil y algunos cánceres más, no se registra ningún otro cambio repentino, en las tasas de muerte por alguno de los cánceres principales, achacable al tratamiento por quimioterapia. Por tanto, para las personas de edad media y madura, el cuadro refleja más una estabilización que un cambio.

Quienes organizan centros de lucha contra el cáncer y supervisan las numerosas pruebas clínicas de la quimioterapia, intentan silenciar o esquivar estas implacables estadísticas. Rechazan a veces la inamovibilidad de las cifras de mortalidad nacional argumentando que inevitablemente van varios años retrasadas, y por tanto, no reflejan los avances más recientes en el tratamiento. Aunque ello es absolutamente correcto, se ha vuelto a plantear reiteradamente la objeción durante los últimos 10 años, sin que lo hayan justificado las estadísticas cuando por fin salían a la luz. Sin embargo, en la mayoría de los casos, los organizadores no toman en consideración las cifras de mortalidad, y se limitan a afirmar que la proporción de pacientes que siguen vivos al cabo de cinco años de emitido el diagnóstico ha aumentado gradualmente en casi todas las clases de cáncer. Atribuyen ese incremento de la supervivencia de cinco años a una progresiva mejora de los métodos de tratamiento.

Antes de negar la fiabilidad de las estadísticas sobre mortalidad, debería considerarse si se comete algún error sistemático en la elaboración de las tasas de supervivencia. Como se ha visto, cabe traducir los resultados de pruebas clínicas a números de vidas salvadas, en toda la nación, cuando se refieren a cánceres como la leucemia infantil, en los que el diagnóstico no admite duda y el resultado depende por completo del éxito del tratamiento. Pero en el caso de cánceres que no son invariablemente fatales, el cálculo se ve dificultado por la carencia de un mecanismo adecuado para estimar el número de vidas susceptibles de ser salvadas.

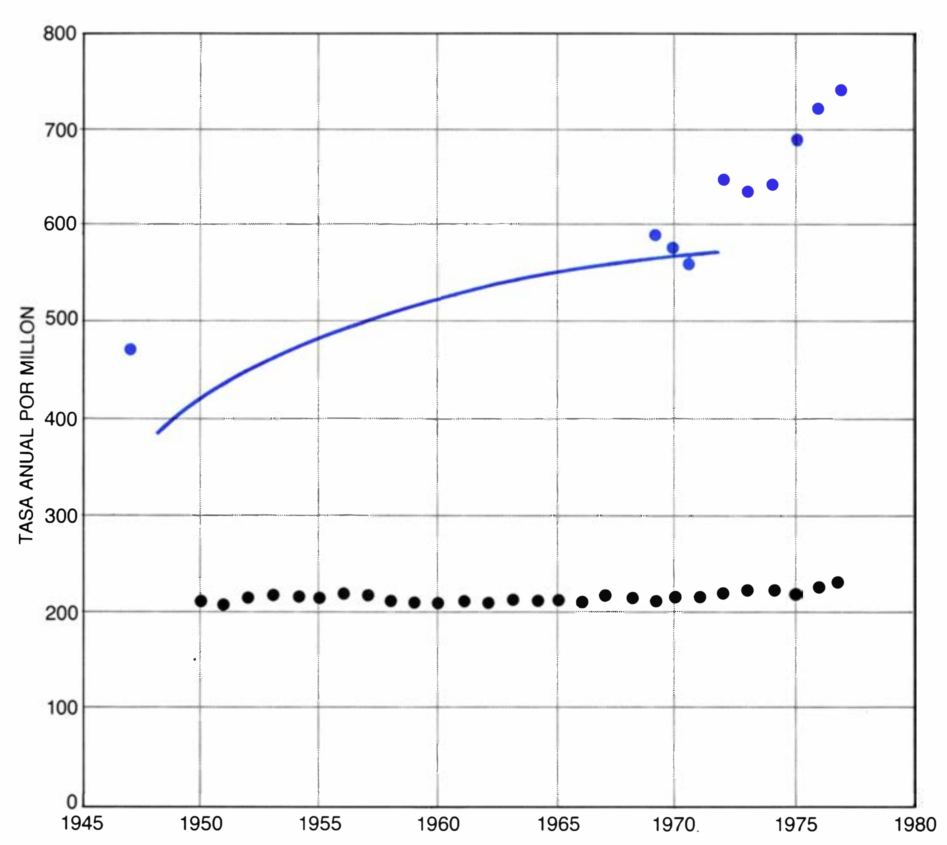
Veamos un ejemplo un tanto extremo. En la cuarta parte de los varones norteamericanos con más de 70 años de edad que murieron por causas distintas del cáncer, se detectaron pequeños cánceres de próstata al someterlos a un examen rutinario postmortem. Sin embargo, sabemos, a partir de los datos de incidencia, que menos del 10 por ciento de esos cánceres estaban destinados a producir síntomas, y una proporción aún menor habría resultado fatal. Por tanto, cualquier campaña encaminada a detectar y tratar cánceres de próstata cuando están todavía en ciernes y antes de que produzcan síntomas incluirá con certeza muchos "cánceres" que no se hubieran detectado de no ser por la propia campaña. Aun cuando la campaña no salve ninguna vida, la inclusión de esos casos adicionales de "cáncer" no fatales incrementaría la proporción de "pacientes" que sobrevivieron.

Algo parecido ha sucedido en Estados Unidos a lo largo de los últimos 30 años. Aunque las pruebas clínicas no han demostrado ningún avance notable en el tratamiento del cáncer de próstata desde la introducción de la terapia con hormonas en la década de 1940, se afirma que la supervivencia de cinco años, con respecto al número de casos, ha pasado del 43 al 63 por ciento. Las estadísticas nacionales muestran, sin embargo, que es la incidencia de casos registrados lo que ha cambiado, pasando de 400 por millón de hombres y año, a finales de la década de 1940, a unos 700 por millón y año a finales de los 70. La tasa de muerte ha permanecido estacionaria, en torno a las 210 fallecimientos por millón de hombres y año. Ha crecido, pues, la tasa de supervivencia, pero no porque mueran menos hombres afectados de cáncer de próstata, sino porque han entrado más pacientes en esa categoría.

Artefactos similares afectan probablemente también a las tasas de supervivencia relacionadas con otros tipos de cáncer; el de mama, en particular. Por ello, constituye hoy un principio fundamental, al menos entre muchos epidemiólogos, el que la comparación de la supervivencia de pacientes en diferentes épocas no suele ser un método aceptable para medir el éxito de la terapéutica (a no ser, como en el caso de la leucemia infantil, que quede bien claro que no ha habido ningún cambio en la definición de lo que constituye la enfermedad). Para la mayoría de las formas de cáncer, los médicos se ven forzados a volver de nuevo a la "infinita aritmética" de los ensayos clínicos. Deben separarse al azar grupos de pacientes, administrándoseles los diversos tratamientos rivales que existan. Las tasas medias de supervivencia de los subgrupos mostrarán entonces cuáles son los tratamientos que aportan un beneficio y cuáles causan un daño.

El caso mejor estudiado se refiere al uso de terapias complementarias tras la intervención quirúrgica en el cáncer de mama. Recientemente se han reunido y resumido los resultados procedentes de gran número de ensayos. En conjunto, las pruebas afectaron a unas 5000 mujeres, sometidas a diversos agentes quimioterapéuticos tóxicos. Un número igual de mujeres no recibió ningún tratamiento adicional después de la cirugía. Hubo un seguimiento de esas pacientes, que duró de uno a diez años. Durante ese período, en el conjunto del grupo tratado se han registrado alrededor de un 25 por ciento menos de muertes que en el grupo control; para las mujeres con edad inferior a 50 años, la reducción ha sido de aproximadamente una tercera parte. No se sabrá si esas pacientes se han curado gracias al tratamiento hasta que el control de los dos grupos se prolongue bastantes años más, pero aunque sólo se haya pospuesto la muerte, nos hallamos ante un claro avance. En cualquier caso, la quimioterapia del cáncer de mama puede ofrecer beneficios reales, por mucho que de momento resulten bastante modestos y no puedan compararse con los obtenidos en ciertos cánceres infantiles.

Ante esos hechos, resultaría tentador concluir que debería someterse a quimioterapia a toda mujer afectada de cáncer de mama, pensando que, después de todo, una tercera parte menos de muertes entre mujeres de edad inferior a los 55 años podría traducirse en una cifra de 2000 o 3000 vidas salvadas cada año, y eso sólo en Estados Unidos. La historia real, sin embargo, es desgraciadamente más complicada de lo que sugieren las estadísticas. Un tratamiento quimioterapéutico de seis o doce meses de duración no sólo constituye una experiencia bastante desagradable, sino que además, conlleva cierta mortalidad intrínseca. Se sabe, además, que muchas de las drogas que se utilizan son cancerígenas: uno de los efectos a largo plazo de la quimioterapia es que entre el 5 y el 10 por ciento de los pacientes que sobreviven mue-



6. AUMENTA LA SUPERVIVENCIA de los pacientes de cáncer cuando se amplía la definición de la enfermedad de modo que contemple también los casos que no están abocados a un desenlace fatal. En ello parece residir la explicación del proclamado aumento de supervivencia en hombres afectados de cáncer de próstata. Desde finales de la década 1950 hasta últimos de la de 1970, la tasa de mortalidad en Estados Unidos (*puntos negros*) se ha mantenido en unas 210 muertes por millón de hombres al año. En el mismo período, el número de casos declarados se ha elevado de unos 400 por millón y por año a unos 700. (Los puntos de color muestran la incidencia registrada en diversas inspecciones norteamericanas y, la línea de color, el incremento registrado en Connecticut.) El incremento de la incidencia quizás explique ese aumento de la supervivencia a cinco años en hombres con cáncer de próstata, que ha pasado en 30 años del 43 por ciento al 63 por ciento, pues los ensayos no han logrado advertir ningún avance importante en el tratamiento.

ren de leucemia en los diez primeros años subsiguientes. Quizá parezca un riesgo menor para quien sufre un cáncer en estado avanzado y en rápido desarrollo, pero merece ser tenido en consideración en el caso de una mujer con un cáncer de mama pequeño y aparentemente localizado. La probabilidad de que esa mujer muera de cáncer antes del quinto año es sólo del 10 por ciento, aun cuando no reciba ningún tratamiento adicional después de aplicarle la cirugía.

Se ha de prestar atención a otros puntos antes de tomar una decisión. En el mismo conjunto de pruebas figuraba un amplio estudio del tratamiento con el inhibidor de los estrógenos tamoxifén. Este fármaco produjo una reducción significativa en el número de muertes, si bien no tan grande como la debida a la quimioterapia. (Vale la pena recordar que el tamoxifén muestra una eficacia mayor en mujeres de más de 50 años, en tanto que la quimioterapia lo es en pacientes más jóvenes.) Dado que el tamoxifén sólo suele producir efectos secundarios leves, podría abogarse en favor de su uso y preferirlo al de las drogas citotóxicas. La cuestión reside ahora en decidir si la quimioterapia citotóxica tiene algo que ofrecer a los pacientes con cáncer de mama que no pueda conseguirse con el tamoxifén. A juzgar por los datos disponibles, la respuesta más acertada es que pueden beneficiarse de ésta muchas pacientes jóvenes y que, para las pacientes de más de 50 años, la combinación de la quimioterapia con el tamoxifén no parece producir mejores resultados que el tamoxifén solo.

Está mucho menos documentado el papel de la quimioterapia en el tratamiento de otros importantes cánceres de adultos. Los resultados de diversas pruebas sugieren que, en los adultos, algunos cánceres, el de ovario entre ellos, responden en ocasiones a la quimioterapia. Por otra parte, puede administrarse quimioterapia e irradiación local para reducir el tamaño de los cánceres de localización inaccesible, verbigracia, ciertas regiones de la cabeza y del cuello.

En términos generales, sin embargo, en cuanto a la duración de la supervivencia se refiere, los resultados han sido más negativos que positivos. Un reciente informe describía una prueba en la que se utilizaba la quimioterapia para el tratamiento del cáncer de colon. A más de 600 pacientes, sometidos previamente a la práctica quirúrgica convencional, se les adjudicó al azar una de las diversas formas de terapia adyuvante. Aproximadamente la mitad de los pacientes recibieron quimioterapia citotóxica (fluoruracilo y un agente alquilante), pero su supervivencia no pudo distinguirse de la de los controles, que no recibieron ningún tratamiento adicional. De los 190 pacientes sujetos a quimioterapia que no murieron a causa de su cáncer durante el período de seis años que duró la prueba, uno falleció de las consecuencias inmediatas del tratamiento y, siete, de leucemia.

A pesar de esos hallazgos más bien modestos, se usan ya comúnmente varias drogas citotóxicas. En el Registro del Cáncer de Connecticut, por ejemplo, consta que alrededor de una cuarta parte del total de los pacientes con cáncer han recibido quimioterapia durante su primera estancia en el hospital. El Instituto Nacional del Cáncer de los Estados Unidos calcula que en ese país se someten a quimioterapia más de 200.000 pacientes cada año. Por ser una forma de tratamiento peligrosa y tecnológicamente exigente, las cifras resultan inquietantes, pues todavía no se sabe con certeza qué beneficio reportan en la mayoría de las categorías de pacientes. Además, el número de pacientes que se curan apenas alcanza un pequeño porcentaje total de enfermos tratados.

A pesar de todo lo dicho, los tratamientos adyuvantes o complementarios evitan algunos miles (tal vez el 2 o 3 por ciento) de las 400.000 muertes por cáncer que se registran cada año en Estados Unidos. Aun cuando no se encontrase ninguna otra droga adicional, esta cifra podría elevarse hasta el 5 por ciento; un uno por ciento más procedería, por ejemplo, del tratamiento del cáncer de mama con agentes citotóxicos. Estas ganancias son patentes, al tiempo que constituyen un monumento funerario dedicado a los miles de pacientes que participaron en las primeras pruebas de la quimioterapia.

El valor y el altruismo de esos pacientes no se ven correspondidos, sin embargo, por el sentido de la responsabilidad que debiera guiar a los encargados de la política sanitaria. En la década de 1960 se estableció que el tabaquismo constituía una causa importante de cáncer de pulmón. ¿Cuántas naciones se han esforzado en contener la expansión de la industria del tabaco desde entonces? Antes bien, existen enormes incentivos económicos para las naciones que dan la espalda al problema. El cigarrillo es un producto sobre el cual se pueden gravar impuestos fácilmente; en Estados Unidos ingresan unos 6000 millones de dólares anuales en las arcas federales y estatales y, lo que es más importante (al menos para el gobierno británico y también a los ojos del gobierno de Estados Unidos), el hábito de fumar recorta el presupuesto destinado a los pensionistas, porque reduce la duración de la vida.

Por el precio de un ligero incremento de los costos de asistencia sanitaria, los fumadores estadounidenses actuales ahorrarán a la Seguridad Social una media de unos 35.000 dólares cada uno: morirán, por término medio, antes que los no fumadores; la mayoría de las muertes ocurren después de la jubilación y no son por cáncer, sino debidas a alguna enfermedad cardiovascular o pulmonar crónica, cuya incidencia eleva también el hábito de fumar. El acortamiento de la vida representará un ahorro total de unos 10.000 millones de dólares al año en el medio siglo próximo.

Algunos países han prohibido toda publicidad del tabaco, y ello ha tenido un efecto casi instantáneo sobre su venta. La incapacidad del gobierno de Estados Unidos para dar ese paso sobrepasa y supera todos los avances realizados en el tratamiento del cáncer desde el advenimiento de la moderna cirugía. A partir de 1953, el cáncer de pulmón constituye la enfermedad fatal más común entre los varones, y se espera que supere al cáncer de mama y sea también la enfermedad mortal más común entre las mujeres. El gasto de vida es realmente asombroso. Gracias al cigarrillo, Estados Unidos registra ahora cada año 100.000 muertes adicionales, a resultas del cáncer de pulmón. Esas cifras dejan en nada las 5000 o 10.000 vidas que salva la quimioterapia. Por el momento, la guerra contra el cáncer está perdida porque (siguiendo con la metáfora) continuamos tolerando la presencia de una quinta columna entre nosotros.

La derrota del más común de los cánceres letales depende, por tanto, de los gobernantes y no de la pericia de los médicos o del ingenio de los científicos. Afortunadamente, en estratos más altos fuman ahora menos de lo que se acostumbraba. Puesto que suelen marcar la pauta, es de esperar que la tendencia se extienda al resto de la población.

En otros cánceres importantes la cuestión no está tan clara. En orden decreciente de importancia numérica figuran el cáncer de intestino grueso, de mama, de próstata y de páncreas. Se trata de afecciones habituales en unos países y raras en otros; cada uno responderá, pues, a causas externas que dominan en algunas partes del mundo y son raras, o no se dan, en otras. Incluso en el propio occidente se distinguen grupos de gente en los que la tasa de muerte por cáncer es sólo la mitad de la tasa media general. Ello prueba que la mayoría de las formas de cáncer puede prevenirse.

No es esta una conclusión sorprendente. Ninguna de las causas importantes de muerte ha podido controlarse fundamentalmente gracias a la terapia. La mortalidad provocada por la malaria, cólera, fiebre tifoidea, tuberculosis, escorbuto, pelagra y otros azotes del pasado ha disminuido gracias, sobre todo, a que la especie humana ha logrado prevenirlas; mas no porque se conozca su tratamiento. Incluso la mortalidad cardiovascular (la forma más habitual de muerte en los países desarrollados) ha empezado a declinar en Estados Unidos, señal de que algún cambio de las circunstancias o del estilo de vida tiende a prevenir su incidencia. Existen, por tanto, motivos para creer que, cuando a una enfermedad se la asedia a gran escala, el principal esfuerzo debe centrarse en evitar su aparición. Dedicar el mayor esfuerzo al tratamiento de la enfermedad es negar todo lo que antecede.

Los cánceres de cérvix y de hígado suelen responder a la infección por parte de algún agente vírico, y deberían poder evitarse por inmunización. Se salvarían con ello unas 500.000 vidas al año en todo el mundo. Las causas de la mayoría de los demás cánceres no se conocen todavía suficientemente bien para predecir cómo ni cuándo podrán prevenirse; es de esperar que algún día se descubran.

No son tan claras las perspectivas de grandes avances en el tratamiento del cáncer. Las drogas citotóxicas disponibles carecen de virtualidad discriminatoria en su acción: son tóxicas para cualquier célula que se divida rápidamente. A primera vista incluso resulta incluso sorprendente que la quimioterapia pueda tener algún éxito. Recuérdese, sin embargo, que los cánceres mejor tratados por quimioterapia se engloban en dos clases especiales. La primera comprende los cánceres derivados de células separadas del proceso de embriogenésis (los cánceres especiales de la infancia), de células de la línea germinal (ciertos cánceres testiculares y ováricos) y de células fetales que quedan atrapadas en la madre (coriocarcinoma de placenta). Estos cánceres comparten una característica poco habitual: son células que se encuentran en un ambiente foráneo. Es bastante probable que el cuerpo posea mecanismos para destruir tales restos, especialmente si la quimioterapia ha reducido su número.

La única neoplasia de fácil curación por quimioterapia es la enfermedad de Hodgkin. Se trata de un cáncer rarísimo, formado por una mezcla de varios tipos de células; de hecho, hasta hace poco se la consideraba una suerte de infección crónica. En pocas palabras, la extrema peculiaridad de esta relación de cánceres sugiere que las formas actuales de quimioterapia sólo raramente, en el mejor de los casos, resultan suficientemente específicas para matar todas las células de un cáncer y no afectar a los tejidos normales del paciente. No se ha demostrado aún que algún cáncer pueda curarse por quimioterapia.

¿Qué decir entonces de las perspectivas a largo plazo? Nadie puede aventurar qué formas nuevas de quimioterapia aparecerán, ni cuándo se inventarán. Mientras se esperan esos descubrimientos resulta preciso dirigir un mayor esfuerzo hacia ciertas formas comprobadas de detección de pacientes con cáncer y concentrar mucha más energía en su prevención. No parece buena política económica la de subvencionar la quimioterapia de los cánceres comunes de adultos y no hacerlo con la exploración preventiva de cáncer de mama. Y lo que es mucho peor, un verdadero acto de locura, es gastar cientos de millones de dólares cada año para administrar quimioterapia a un número creciente de pacientes, mientras nada se hace para proteger a la población del peligro del tabaquismo.

Rayos cósmicos de Cisne X-3

Tras decenios de búsquedas infructuosas, los astrónomos han encontrado una fuente de partículas de alta energía y de rayos gamma que bombardean la Tierra. Se trata de una estrella binaria situada en el borde de la galaxia

P. Kevin MacKeown y Trevor C. Weekes

Pocas cuestiones han tenido perplejos durante tanto tiempo a los astrónomos, desafiando su imaginación, como la del origen de la radiación cósmica. Las propiedades de ésta se han ido conociendo gradualmente, década tras década, desde 1912, año en que el físico austríaco Victor F. Hess demostraba su existencia. Está formada principalmente por núcleos atómicos, carentes por tanto de electrones, que se mueven a velocidad cercana a la de la luz. Los núcleos encierran tanta energía que la potencia disipada en nuestra galaxia en forma de rayos cósmicos se cree que es mucho mayor que la que se radía, por ejemplo, en la banda de rayos X o en ondas de radio. Ahora bien, muchos de los avances más apasionantes de la astronomía han surgido del análisis detallado de las fuentes de rayos X o de ondas de radio, mientras que hasta hace poco la cuestión del origen de los rayos cósmicos permanecía en el terreno de la especulación. Parecen venir de todas partes, cayendo sobre la Tierra desde todas las direcciones a un ritmo uniforme.

Por fin, acaba de hallarse una fuente importante. El objeto se denomina Cisne X-3, porque es el tercer emisor más brillante de rayos X en la constelación del Cisne; los astrónomos lo observaron por primera vez en la banda de rayos X a finales de la década de los 60. Ganó notoriedad en 1972, año en que sufrió una violenta explosión en la que la intensidad de sus emisiones de radio se multiplicó por mil. Recientemente se ha descubierto que constituye también una fuente de rayos gamma de alta energía. Son los rayos gamma los que han permitido identificar a Cisne X-3 como una fuente de radiación cósmica: sólo pueden haberse producido por partículas cargadas que se movieran a una velocidad próxima a la de la luz.

No se dispone todavía de una explicación aceptada por todos en torno a la forma en que las partículas son aceleradas hasta velocidades tan altas; se han ofrecido varios modelos para el proceso. La mayoría de científicos están de acuerdo en que Cisne X-3 es un sistema estelar binario, distante de nosotros 37.000 años luz, por lo menos; se halla, pues, casi en el límite de la galaxia. No puede contemplarse en luz visible, ni siquiera con los mayores telescopios, ya que lo impiden las nubes de polvo de los brazos espirales de la galaxia. Sin embargo, las emisiones observadas a otras longitudes de onda indican que es uno de los dos o tres objetos de la Vía Láctea con mayor luminosidad intrínseca.

Toda fuente de radiación cósmica debe ser un potente acelerador de partículas que justifique la ingente energía de los distintos rayos cósmicos. Los dotados de mayor energía superan los 10^{20} electronvolt. (Es decir, 100 millones de veces más que la energía que se espera suministrar a las partículas en el proyectado Supercolisionador Superconductor, que será el acelerador de partículas más potente hecho por el hombre.) El otro extremo del espectro energético de los rayos cósmicos se define un poco arbitrariamente: cualquier cuanto mayor que 10^{8} electronvolt que llegue del espacio se considera un rayo cósmico. La definición abarca no sólo partículas, sino también fotones gamma, que son cuantos de radiación electromagnética.

Sin embargo, los rayos gamma no alcanzan el 0,1 por ciento del total de rayos cósmicos; el resto son partículas. Los electrones representan en torno al 1 por ciento del flujo total; hay también algunos positrones (antielectrones) y algunos antiprotones, pero la inmensa mayoría de los rayos cósmicos son núcleos atómicos. Aproximadamente el 92 por ciento de ellos son iones de hidrógeno, que no son más que protones individuales. El helio, el segundo elemento más ligero, da cuenta de alrededor del 6 por ciento del flujo. Los otros elementos habituales se encuentran en cantidades mucho menores.

La característica más importante de la composición química de la radiación cósmica reside en su estrecho parecido, en la mayor parte del intervalo de energías observado, con la composición global de la galaxia. Este rasgo, trivial a primera vista, desecha varias hipótesis acerca del origen de los rayos cósmicos. Por ejemplo, no pueden proceder todos de la gran explosión ("big bang"), que produjo casi exclusivamente hidrógeno. Ni pueden venir todos de estrellas viejas, muy evolucionadas, porque entonces cabría esperar que una fracción mucho mayor de rayos fuesen núcleos más pesados. (Se cree que los elementos más pesados que el helio se sintetizaron en el interior de las estrellas y durante las explosiones de las supernovas, mediante reacciones de fusión.) Los rayos cósmicos se originan en un objeto de composición típica o bien deben hacerlo en un gran número de objetos distintos que en conjunto produzcan una mezcla promediada de elementos.

1. SEGUN LOS MODELOS DE CISNE X-3 trataríase de un sistema estelar binario en el que uno de los componentes es una densa estrella de neutrones. De acuerdo con una de las hipótesis, la estrella de neutrones es un púlsar en rápida rotación (*arriba*). El púlsar acelera los protones hasta energías propias de los rayos cósmicos y los expulsa en todas direcciones. Algunos protones chocan contra los núcleos de gas de las capas externas de la estrella acompañante, generando rayos gamma de alta energía en la dirección de la trayectoria del protón. Los rayos gamma se detectan en la Tierra durante dos fases de la órbita del púlsar. En el modelo de acreción (*abajo*), el campo gravitacional de la estrella de neutrones atrae materia de la estrella acompañante y su campo magnético induce un campo eléctrico intenso en el disco de acreción giratorio. Los protones se aceleran hasta alcanzar las energías de los rayos cósmicos a lo largo de las líneas del campo eléctrico; algunos chocan contra núcleos de gas de la corriente de acreción y producen rayos gamma. Los rayos gamma de alta energía de Cisne X-3 le han identificado como una fuente distante unos 37.000 años luz.

PROTONES
PULSAR
RAYO GAMMA
ESTRELLA
ACOMPAÑANTE
DIRECCION
DE LA TIERRA

CHORRO
ESTRELLA
ACOMPAÑANTE
DISCO DE ACRECION
CAMPO ELECTRICO
CAMPO MAGNETICO
CORRIENTE DE ACRECION
ESTRELLA
DE NEUTRONES
PROTON
RAYO GAMMA
DIRECCION
DE LA TIERRA

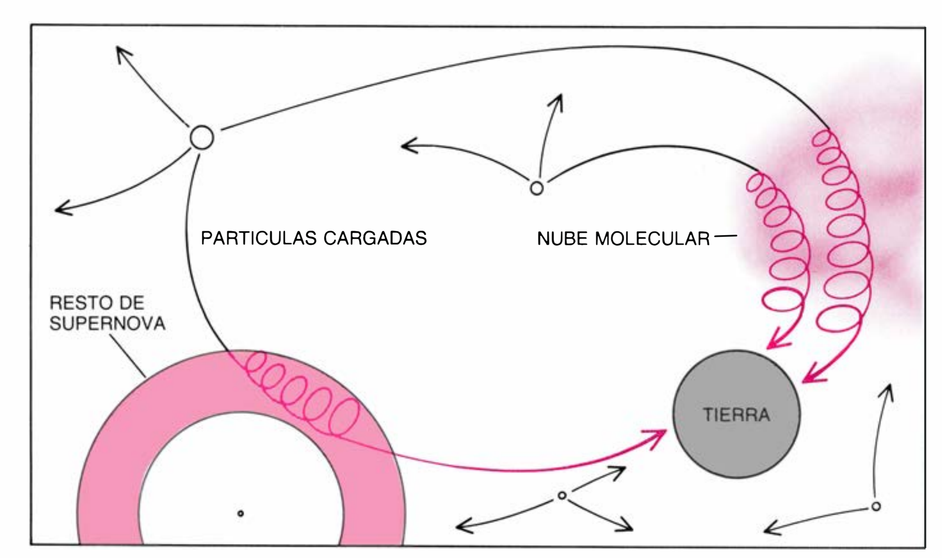
Otras observaciones excluyen la posibilidad de que la radiación cósmica que bombardea hoy día la Tierra pueda ser consecuencia de un solo suceso catastrófico y local; verbigracia: una explosión solar o el estallido de una supernova cercana. En primer lugar, el examen de las trazas de ionización dejadas en meteoritos por partículas cargadas de los rayos cósmicos sugiere que la densidad de los rayos en el sistema solar ha permanecido sensiblemente constante durante los últimos mil millones de años. Las trazas de ionización se advierten en forma de defectos en la estructura cristalina; el número de defectos del meteorito, junto con su edad (que se conoce a partir de la datación radioisotópica), proporciona el ritmo medio con el que el meteorito ha sufrido el bombardeo de los rayos cósmicos.

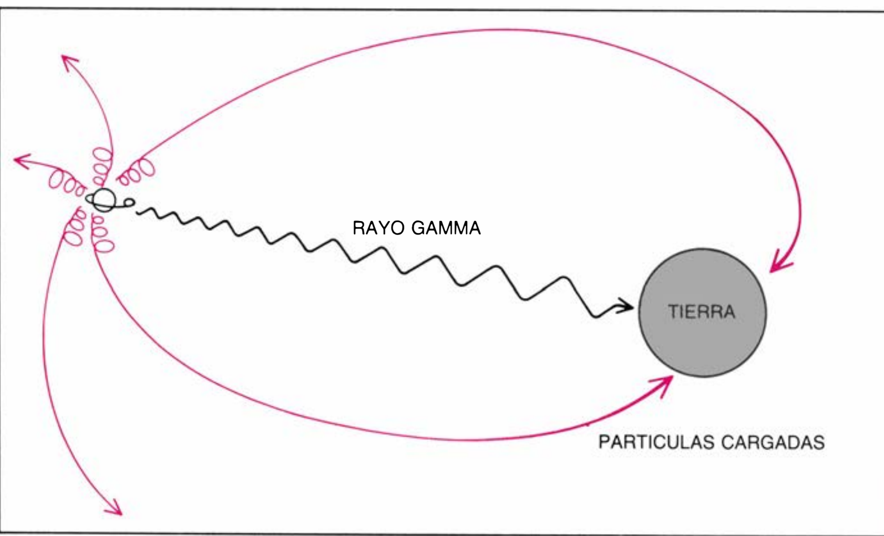
En segundo lugar, la densidad de rayos cósmicos resulta ser la misma en el sistema solar que en el resto de la galaxia. A esta conclusión se llega de un modo indirecto: partiendo de la observación en toda la galaxia de emisiones de sincrotrón en longitudes de radioonda. La radiación de sincrotrón, que tiene una polarización característica, la generan principalmente electrones relativistas que se mueven en un campo magnético. Un electrón relativista en el espacio es un rayo cósmico. Si suponemos que los electrones y las demás partículas de los rayos cósmicos se hallan en las zonas de emisión de sincrotón en la misma proporción que se observa cerca de la Tierra, se puede llegar a una primera estimación de la densidad global de rayos cósmicos en zonas lejanas de la galaxia.

Conocida la distancia que las partículas de alta energía recorren en el espacio antes de chocar con los núcleos del gas interestelar, se puede calcular incluso el ritmo con el que la galaxia debe producir rayos cósmicos para que persista la densidad observada. Las estimaciones de la vida media de los rayos cósmicos se fundan en un examen pormenorizado de la composición de la radiación. Aunque la composición es básicamente similar a la global de la galaxia, algunos elementos ligeros (en concreto, litio, berilio y boro) se encuentran en cantidades mayores que las que serían de esperar. Se cree que estos elementos no se producen por reacciones de fusión en las estrellas. Es casi seguro que resultan de la fragmentación de elementos pesados que han chocado con átomos de gas en reposo en el espacio interestelar. A partir de su abundancia relativa en la radiación cósmica y de una estimación de la densidad de gas interestelar, se calcula el tiempo que, en promedio, los rayos cósmicos han permanecido en el espacio: unos 20 millones de años.

Junto con las energías observadas, esta cifra conduce a una estimación de la potencia disipada por la galaxia en forma de radiación cósmica. Obviamente, un cálculo de este tenor deja abiertos muchos interrogantes, pero se admite que el valor correcto está entre 10^{40} y 10^{41} erg por segundo. La galaxia genera menos energía en forma de rayos cósmicos que la radiada en luz visible (10^{44} erg por segundo) y más de lo que radía en total entre ondas de radio y rayos X (10^{39} y 2×10^{39} erg por segundo respectivamente). Cualquiera que fuere la fuente de los rayos cósmicos, está claro que se producen a un ritmo prodigioso.

A lo largo de la historia han dominado dos interpretaciones principales del origen de la radiación cósmica. De acuerdo con la primera, que propusiera el físico Enrico Fermi, los núcleos de los rayos cósmicos se inyectan en el espacio con energías bastante bajas, para acelerarse luego hasta altas velocidades a través de las colisiones con nubes de gas magnetizado, ondas de choque de explosiones de supernovas u otras configuraciones energéticas de gran tamaño del medio interestelar. Se sabe que los núcleos de baja energía son arrojados al espacio por las estrellas ordinarias; según el modelo de





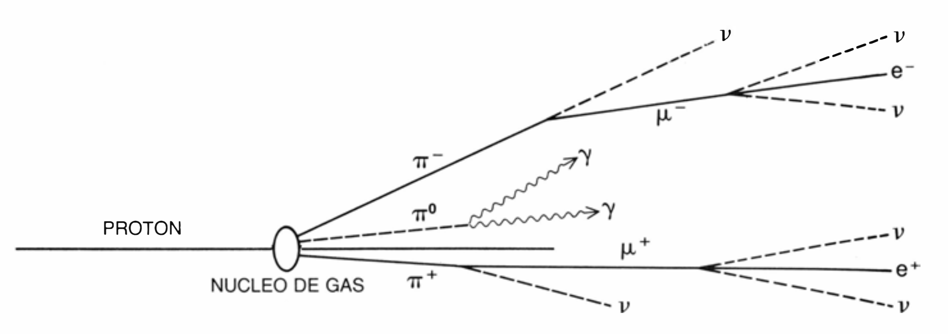
2. ORIGEN DE LOS RAYOS COSMICOS y su doble explicación. Según el modelo de Fermi (*esquema superior*), los rayos cósmicos se producen en el espacio interestelar. Las partículas de baja energía emitidas por las estrellas ordinarias serían aceleradas hasta las velocidades de los rayos cósmicos al chocar contra nubes de gas magnetizado en movimiento o contra ondas de choque en expansión de explosiones de supernovas. De acuerdo con una segunda hipótesis, los rayos cósmicos surgirían de un pequeño número de objetos exóticos capaces de acelerar partículas a altas energías (*abajo*). Los núcleos cargados que forman la mayor parte de la radiación son desviados por el campo magnético galáctico, de manera que sus direcciones no indican su origen. Los rayos gamma, eléctricamente neutros, viajan en línea recta desde su mismo origen.

Fermi, tales núcleos adquieren las energías propias de rayos cósmicos paso a paso, en un largo recorrido aleatorio a través del espacio. Hace treinta años, este modelo de "aceleración distribuida" gozaba de amplia aceptación, porque evitaba la necesidad de postular la existencia de objetos exóticos y aún no descubiertos capaces de impartir de una sola vez velocidades relativistas a las partículas.

Pero en los últimos años el modelo de Fermi ha caído en desgracia. Cálculos detallados han demostrado que no es probable que las energías de los núcleos aumenten lo suficiente en su camino a través del espacio interestelar para llegar a las observadas en los rayos cósmicos. Además, desde que Fermi propuso el modelo, en astronomía ha sido corriente la observación de objetos exóticos, dotados de extrema energía. Nadie en 1950 podía haber previsto el descubrimiento de los púlsares, por ejemplo: estrellas de neutrones densas, en rápida rotación, con campos magnéticos billones de veces más intensos que el terrestre.

Son muchos los investigadores que hoy prefieren una segunda interpretación: los rayos cósmicos reciben toda su energía de fuentes discretas, actuando el espacio interestelar sólo como un medio difusor. Se han propuesto muchos tipos de fuentes, desde los púlsares hasta las explosiones de supernovas, pasando por las estrellas T Tauri (estrellas jóvenes y variables que algunas veces sufren rápidos aumentos de brillo). Algunos de estos objetos se encuentran en número suficiente para justificar la energía total contenida en la radiación cósmica. Sin embargo, parecen incapaces de producir todo el espectro de energías.

Por ello, algunos investigadores sugieren que los rayos cósmicos, o por lo menos los más energéticos, deben proceder de allende nuestra galaxia. La emisión de sincrotrón de otras galaxias muestra que también ellas poseen rayos cósmicos. Ciertas galaxias tienen núcleos activos violentos, algunos de los cuales expelen partículas en forma de chorros relativistas; se supone que los quásars serían galaxias extremadamente activas. Si una galaxia posee un núcleo activo, es concebible que pueda acelerar partículas a energías que están fuera del alcance de una estrella, individual o binaria, de nuestra propia galaxia. Por otro lado, esto no implica que todos los rayos cósmicos deban originarse fuera de la Vía Láctea, según han sugerido otros investigadores.



3. SE SUPONE QUE LOS RAYOS GAMMA DE CISNE X-3 son producidos por el choque de protones de alta energía procedentes de la estrella de neutrones con núcleos de gas de la estrella acompañante. Parte de la energía cinética del protrón se convierte en partículas inestables llamadas piones. Cada pion cargado produce un neutrino y un muon; éste se desintegra a su vez en dos neutrinos más y en un electrón o un positrón. El pion neutro se desintegra en dos rayos gamma altamente energéticos. (Dibujo de Andrew Christie.)

Podría pensarse que las fuentes de los rayos cósmicos dejarían identificarse tras unas cuantas observaciones sencillas con un telescopio de rayos cósmicos. No es así. Los núcleos de los rayos cósmicos vienen de todas direcciones y casi exactamente con el mismo ritmo, pero no porque las fuentes estén distribuidas uniformemente, sino porque las trayectorias de las partículas están revueltas a causa del campo magnético de la galaxia. Por hallarse eléctricamente cargados, los núcleos no viajan en línea recta a través del espacio interestelar. Antes bien, se ven obligados a describir espirales a lo largo de las líneas del campo magnético, acerca de cuya configuración, desordenada y caótica, poco se sabe. Cuando las partículas llegan a la Tierra después de 20 millones de años de carrera por la galaxia, sus trayectorias están intrincadamente enmarañadas.

Para localizar una fuente, hay que estudiar componentes de la radiación que sean relativamente insensibles a los campos magnéticos. Los candidatos más claros son las partículas eléctricamente neutras. De los neutrones podemos olvidarnos, ya que se desintegrarían en partículas cargadas (protones y electrones) mucho antes de que pudieran alcanzar nuestro planeta arrancando de una fuente de rayos cósmicos. Mayores esperanzas se han puesto en los neutrinos: seguramente se generarán en las reacciones nucleares que se desarrollen en la fuente de rayos cósmicos, y es bien conocida su capacidad de atravesar obstáculos sin quedar absorbidos. Por desgracia esta misma propiedad los convierte en muy difíciles de detectar. No se ha construido todavía ningún telescopio capaz de detectar neutrinos de alta energía provenientes de fuentes cósmicas, aunque los hay en fase de proyecto.

El planteamiento que nosotros hicimos, compartido por otros investigadores, ha dado su fruto al confirmarse la detección de radiación cósmica de Cisne X-3. Tratábase de observar fotones gamma que, como los neutrones y los neutrinos, son eléctricamente neutros. Es casi inevitable que un objeto que acelere partículas a las velocidades propias de los rayos cósmicos produzca también rayos gamma; y a la inversa: los rayos gamma de alta energía sólo se emiten por partículas relativistas. Los rayos gamma constituyen, por tanto, una sonda eficaz para buscar fuentes de radiación cósmica, a pesar de que representan menos del 0,1 por ciento del flujo total. Además, en la mayoría de longitudes de onda, interaccionan muy poco con el tenue gas interestelar; y llegan, pues, al sistema solar casi sin sufrir atenuación alguna.

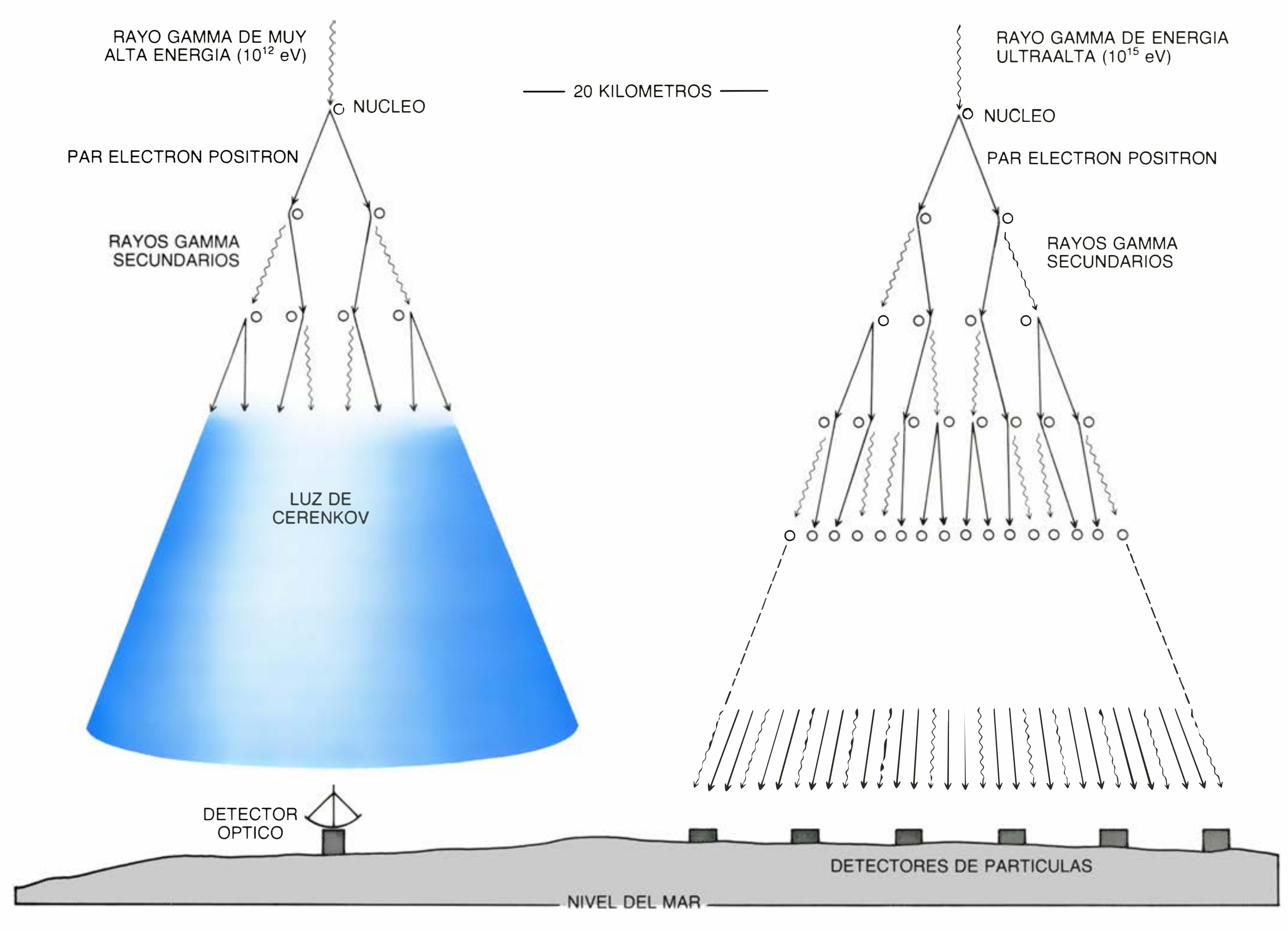
Los rayos gamma cósmicos no alcanzan, sin embargo, la superficie de la Tierra. Antes de que hayan penetrado más de un cuarentavo de la densa atmósfera interaccionan con los núcleos de gas. Por tanto, sólo pueden detectarse directamente con instrumentos situados sobre la atmósfera. Añádase a ello que los instrumentos de los satélites tienen un tamaño limitado, lo que les incapacita para detectar flujos menores de un fotón por metro cuadrado al mes. El flujo de los rayos gamma de energías por encima de 10^{12} electronvolt, cuyo origen es el más intrigante, está por debajo de dicho límite. No pueden estudiarse directamente.

Por fortuna para nosotros, estos rayos gamma sí pueden estudiarse indirectamente desde el suelo, porque provocan en la atmósfera cascadas observables de partículas secundarias y de fotones. Una cascada atmosférica (conocida también por cascada electromagnética) se desencadena cuando un

rayo gamma cósmico interactúa con el campo eléctrico de un núcleo de gas en la alta atmósfera, a una altitud típica de 20 kilómetros. La energía del rayo gamma se transforma en materia: normalmente en un electrón y un positrón, cada uno de los cuales viene a llevarse la mitad de la energía del rayo gamma. Tras un corto recorrido, cada una de estas dos partículas de alta velocidad es desviada, a su vez, por el campo eléctrico de un núcleo del gas y parte de su energía se radía en forma de un fotón gamma mediante un proceso llamado *bremsstrahlung* (radiación de frenado). Estos rayos gamma secundarios producen más pares electrón-positrón. De esta manera, la energía del rayo gamma primario se distribuye entre un número de partículas y fotones que crece en progresión geométrica. La cascada se agota cuando las energías de los cuantos individuales son tan bajas que otros procesos de absorción son de magnitud comparable al *bremsstrahlung* y a la producción de pares.

Si el rayo gamma primario es un fotón de energía ultraalta (10^{14} electronvolt o más), la cascada atmosférica puede alcanzar el suelo. Por entonces se compone ya de miles de partículas y rayos gamma distribuidos en centenares de metros cuadrados. Puede detectarse con una red de contadores Geiger (en los que las partículas cargadas ionizan moléculas de gas, lo que permite que una corriente fluya entre dos electrodos) o contadores de centelleo (en los que los destellos de luz provocados por las partículas en un medio adecuado son registrados por tubos fotomultiplicadores). Cuando la dirección de llegada del rayo gamma primario es vertical, las partículas secundarias inciden, simultáneamente, en todos los detectores de la red. Si el rayo gamma primario incide en la atmósfera con un cierto ángulo, se inferirá su dirección de llegada a partir de las pequeñas diferencias, típicamente inferiores a una diezmillonésima de segundo, entre los tiempos de llegada de las partículas de la cascada a los distintos detectores. La energía de los rayos gamma primarios puede estimarse conociendo el número de partículas que arriban a los detectores.

La cascada atmosférica provocada por un rayo gamma de muy alta energía (entre 10^{11} y 10^{14} electronvolt) no alcanza la superficie de la Tierra. Sin embargo, antes de que las partículas de la cascada sean absorbidas por la atmósfera, desencadenan destellos de luz visible azulada que pueden detectarse desde el suelo. La luz, llamada radiación de Cerenkov, se genera cuando una partícula se mueve en un medio a una velocidad mayor que la velocidad de la luz en ese medio; de hecho la emiten los electrones del propio medio al ser perturbados por el paso de la partícula. La radiación de Cerenkov es un análogo electromagnético de la ola arqueada que genera una barca potente al moverse más deprisa que la velocidad de propagación de las ondas en el



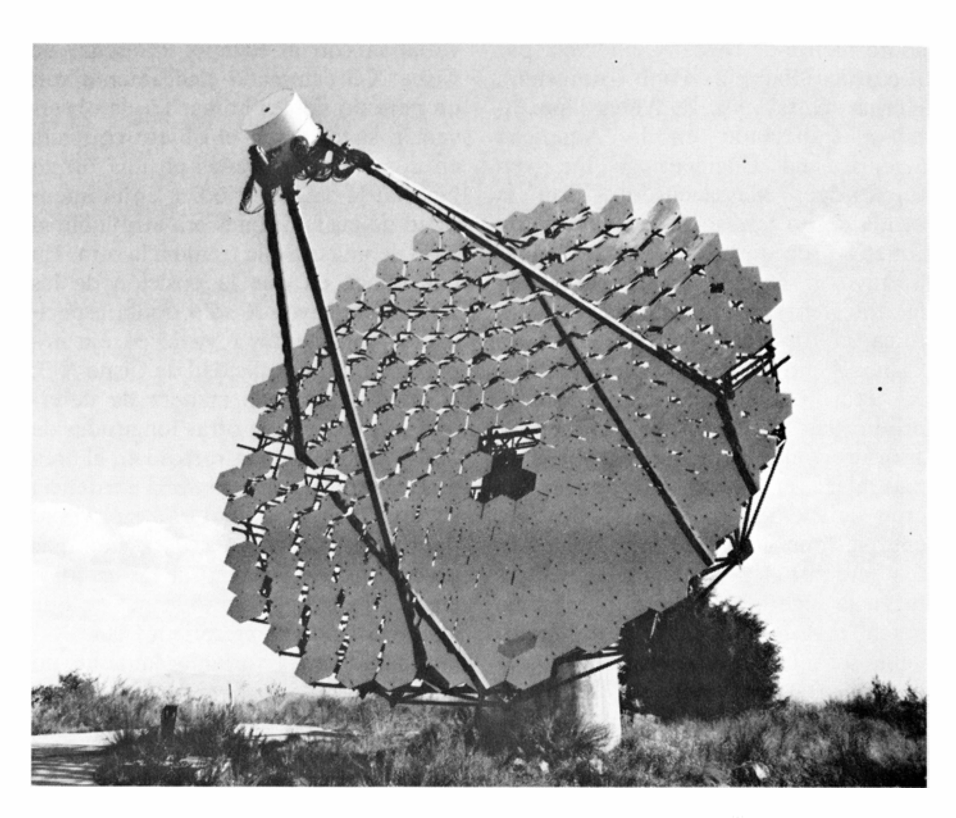
4. CASCADAS ATMOSFERICAS provocadas por los rayos gamma cósmicos. Nos permiten determinar la dirección de llegada de los rayos gamma. Estos no suelen alcanzar la superficie de la Tierra, pues interaccionan con núcleos atmosféricos a unos 20 kilómetros de altura. La energía del rayo gamma se transforma en un electrón y un positrón. Cada partícula emite un rayo gamma secundario al sufrir una desviación por parte del campo eléctrico de otro núcleo. El rayo secundario da lugar así a otro par electrón-positrón y el proceso continúa hasta que se disipa la energía del rayo primario. Una cascada atmosférica provocada por un rayo gamma de energía ultraalta puede llegar al suelo, donde las partículas se observarán mediante contadores Geiger u otros detectores. La cascada atmosférica de un rayo gamma de muy alta energía se absorbe en la atmósfera, pero algunas partículas secundarias generan luz de Cerenkov.

agua. Está dirigida según la trayectoria de la partícula y llega al suelo en forma de disco de luz. Como la luz está en el dominio del visible, sufre sólo una leve atenuación en la atmósfera.

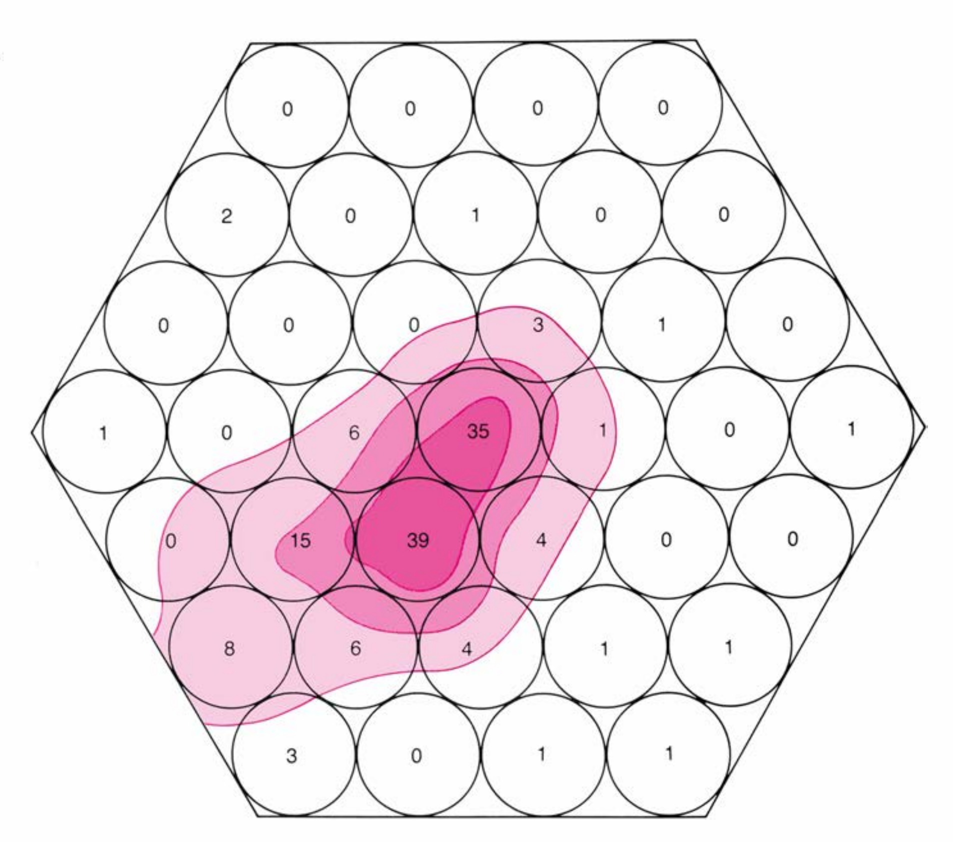
En las noches sin luna, los destellos de la luz de Cerenkov se detectan con un dispositivo sencillo y barato: un espejo que focaliza la luz en un fototubo. Dispositivos más refinados, como el que se usa en el Fred Lawrence Whipple Observatory, en Arizona, constan de muchos espejos reunidos en una plataforma que focalizan la luz sobre una serie de fototubos. Tales dispositivos no sólo detectan luz de Cerenkov, sino que aportan también su diagrama de intensidad. La dirección de llegada de los rayos gamma cósmicos, que corresponde aproximadamente a la del eje del disco de la luz de Cerenkov, puede determinarse entonces con un error menor que un cuarto de grado de arco. Por contra, la resolución de las configuraciones corrientes de detectores de partículas es de sólo un grado de arco aproximadamente. Ahora bien, los detectores de partículas pueden utilizarse las 24 horas del día y no sólo en las noches oscuras; ello les convierte en el principal método de observación de los rayos gamma cósmicos de energías de 10^{15} electronvolt o más.

Ambos tipos de dispositivos tienen un punto débil importante: no permiten asegurar que una cascada atmosférica determinada la haya desencadenado un rayo gamma. El problema estriba en que las partículas cargadas de los rayos cósmicos interactúan con los núcleos atmosféricos provocando cascadas similares a las causadas por los rayos gamma. Además, como resultado de la influencia del campo magnético galáctico, el flujo de rayos cósmicos cargados viene a ser el mismo en todas direcciones. Por tanto, una fuente de rayos gamma cósmicos puede manifestarse como una anisotropía direccional: un exceso de cascadas atmosféricas procedentes de una dirección determinada. La anisotropía será necesariamente pequeña, ya que las cascadas atmosféricas provocadas por los rayos cósmicos son aproximadamente mil veces más numerosas que las desatadas por los rayos gamma. Esta es una de las razones por las que se ha invertido años de laboriosas observaciones y análisis de datos para identificar una fuente de rayos cósmicos.

La historia de Cisne X-3 empezó mucho antes del reciente reconocimiento de su importancia como fuente de rayos cósmicos. Fue descubierta



5. TELESCOPIO DE LUZ DE CERENKOV, utilizado por los autores y sus colaboradores en el Observatorio Fred Lawrence Whipple, Arizona; se le conoce también por reflector de 10 metros. Consta de 248 espejos que focalizan la luz en 37 fototubos dispuestos de forma hexagonal. En las noches oscuras, el instrumento puede detectar fácilmente los destellos de luz de Cerenkov emitidos por las cascadas atmosféricas.



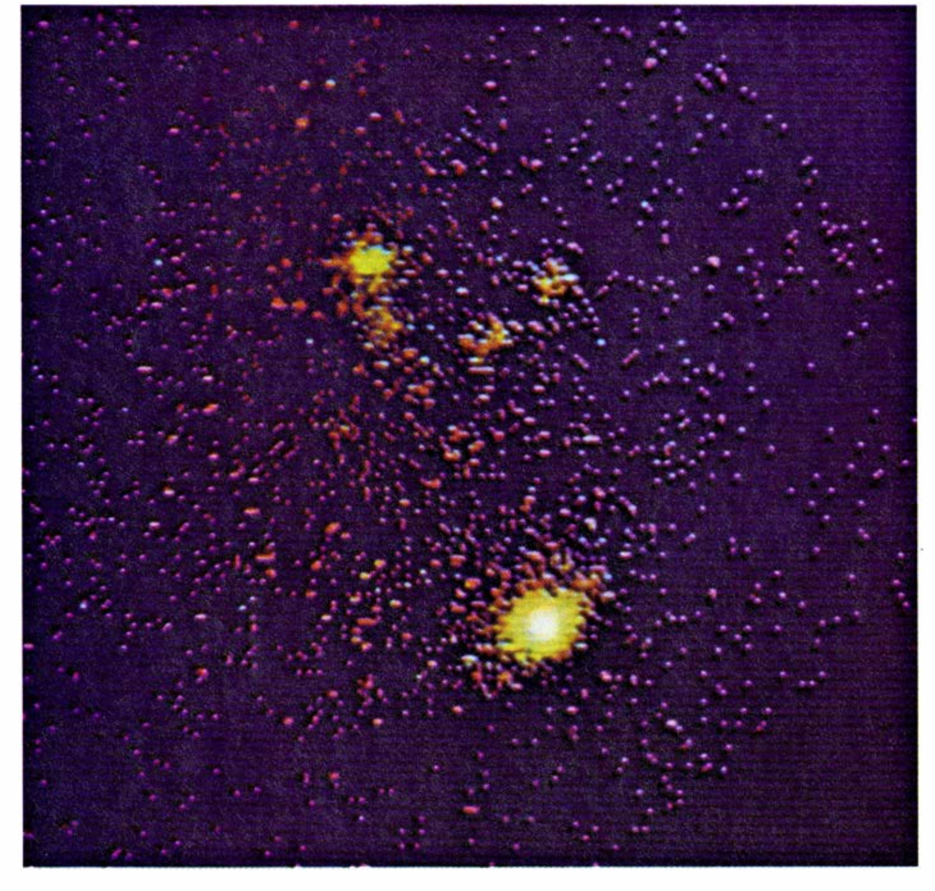
6. MAPA DE LUZ DE CERENKOV de una cascada atmosférica, trazado con el reflector de 10 metros en el Observatorio Whipple. Los círculos representan los fototubos. Los números son proporcionales a la intensidad de la luz registrada por cada fototubo; las líneas coloreadas definen contornos de intensidad. A partir de un mapa como éste puede calcularse el eje de la cascada atmosférica y con ello la dirección de llegada del rayo cósmico. En nuestro caso, el eje de la cascada corre paralelo al eje del detector y está desplazado de éste hacia su parte inferior izquierda. (El mapa sólo registra una parte del borde de la cascada cónica.) Los rayos gamma cósmicos de Cisne X-3 se detectan a través del exceso de cascadas en dicha dirección.

como fuente de rayos X, en 1966, por Riccardo Giacconi, Paul Gorenstein, Herbert Gursky y J. R. Waters, que estaban trabajando en la American Science and Engineering, Inc., en Cambridge, Massachusetts. Con la ayuda de un telescopio a bordo de un cohete (procedimiento necesario, pues los rayos X son también absorbidos por la atmósfera), los investigadores identificaron tres fuentes de rayos X en Cisne. Aunque las emisiones de Cisne X-3 eran bastante más débiles que las producidas por las otras dos fuentes, aportaron una pista que llevaba a su naturaleza energética. Las emisiones eran deficientes en rayos X de baja energía, que son absorbidos preferentemente por el gas interestelar. La deficiencia sugería que Cisne X-3 distaba mucho más de lo supuesto y, por tanto, debía ser intrínsecamente más brillante que las otras dos fuentes.

Cuando la Administración Nacional de la Aeronáutica y el Espacio (NASA) lanzó su primer telescopio de rayos X a bordo de un satélite (*Uhuru*), en 1970, se descubrió que muchas señales de rayos X, incluida la del Cisne X-3, variaban con el tiempo. La señal de Cisne X-3 cambiaba cíclicamente con un período de 4,8 horas. La de observación sugería que el objeto consistía en dos estrellas ligadas en una órbita binaria; la caída periódica en la intensidad de las emisiones era atribuible al paso de una estrella frente a la otra. En una época en que la posición de las fuentes de rayos X sólo podía especificarse vagamente, revistió pareja importancia la periodicidad de Cisne X-3, ya que ofrecía una manera de determinar si radiaba en otras longitudes de onda. Estudios de infrarrojo en el área de Cisne revelaron pronto la existencia de una fuente puntual con su propio período de 4,8 horas. La posición más precisa proporcionada por los estudios en el infrarrojo permitió, a su vez, que los investigadores viesen en Cisne X-3 una radiofuente variable (aunque no periódica).

Así las cosas, el 2 de septiembre de 1972 un descubrimiento fortuito acabó con la relativa oscuridad de Cisne X-3, mostrando cuán drásticas eran las variaciones de sus radioemisiones. A primera hora del atardecer, mientras esperaba que su principal objetivo se levantara sobre el horizonte, Philip C. Gregory, que trabajaba en el Algonquin Radio Observatory de Ontario, decidió pasar el tiempo apuntando su radiotelescopio a Cisne X-3. La que ordinariamente parecía una radiofuente débil, aquella noche brillaba con un esplendor singular, radiando a una intensidad de mil veces la normal. Un fenómeno equivalente para un astrónomo óptico sería el de una estrella débil que adquiriera, de repente, la luminosidad de Júpiter.

Gregory telefoneó inmediatamente a su colega Robert M. Hjellming, al National Radio Astronomy Observatory en Green Banck, West Virginia. Hjellming confirmó la observación. Aquella misma noche, temiendo que la explosión pudiera acabar antes de lograr registrarla adecuadamente, los dos investigadores llamaron a cuantos astrónomos pudieron. En pocos días, las noticias de la explosión de Cisne X-3 habían llegado a observatorios de todo el mundo. Se interrumpieron los programas regulares, y telescopios de todo tipo –de radio, de infrarrojo, ópticos, de rayos X y de rayos gamma– se apuntaron hacia Cisne. Nunca hasta entonces se había producido una concentración tal de esfuerzos en un único objeto.



7. IMAGEN DE RAYOS X DE CISNE X-3. Fue preparada por Rick F. Harnden, Jr., del Observatorio del Harvard College y del Smithsonian a partir de los datos obtenidos por el Observatorio Einstein en órbita. Cisne X-3, descubierto como emisor de rayos X, constituye la fuente mayor y más brillante de la imagen; se trata de una de las fuentes más luminosas de la galaxia. Las otras fuentes son todas estrellas jóvenes y masivas.

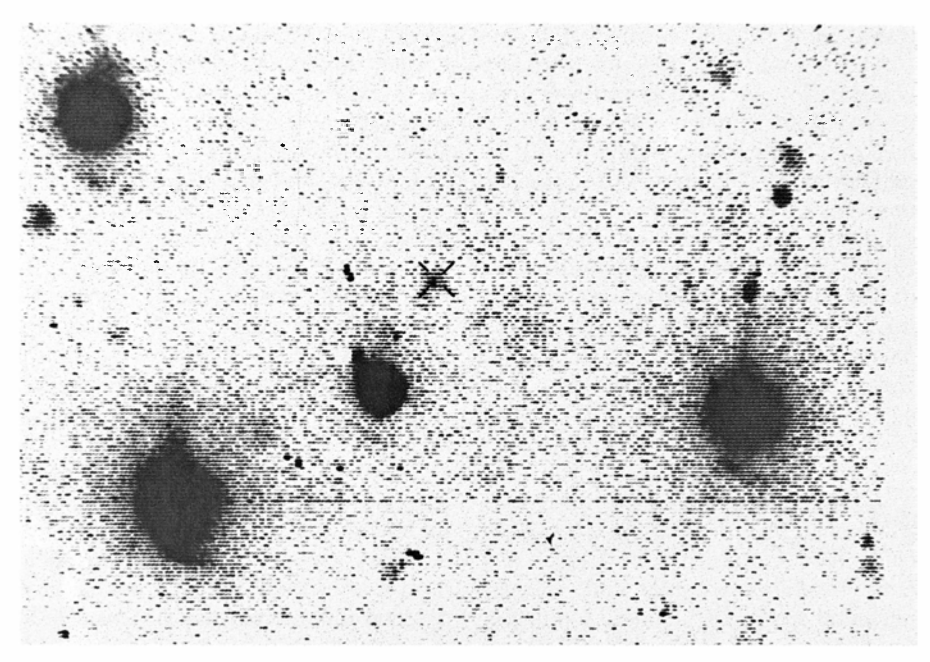
La compleja señal de radio resultó ser rica en información. Por un lado, la señal mostraba líneas de absorción, atribuibles a nubes específicas de hidrógeno interestelar cuyas localizaciones aproximadas en la galaxia eran conocidas. (El hidrógeno absorbe radiación a una longitud de onda de 21 centímetros; por moverse como las nubes interestelares a distintas velocidades, las diferentes nubes ofrecen líneas de absorción características que están desplazadas por efecto Doppler con respecto a la línea de 21 centímetros en cantidades diferentes.) Las líneas de absorción establecieron un límite inferior a la distancia de Cisne X-3: debe estar aún más lejos que la nube de hidrógeno más alejada, que se cree se halla a unos 37.000 años luz, cerca del límite de la galaxia. A partir de la distancia mínima y la luminosidad de rayos X observada de Cisne X-3 podemos calcular su luminosidad intrínseca de rayos X. El cálculo demuestra que se trata de una de las fuentes más brillantes de la galaxia, con una potencia de emisión de por lo menos 2×10^{37} ergs por segundo.

Además, el espectro de las emisiones

de radio observado durante la explosión de 1972 indicaba que Cisne X-3 estaba emitiendo partículas de alta energía. Varios días después del descubrimiento de Gregory, las emisiones de radio alcanzaron un pico de intensidad. Comenzaron entonces a debilitarse de una manera peculiar: las emisiones de longitudes de onda mayores, que correspondían a fotones de menores energías, duraron un tiempo mayor. Este fenómeno es característico de la radiación de sincrotrón de electrones relativistas inyectados en un campo magnético débil. A medida que los electrones describen espirales en torno a las líneas de campo emiten fotones, inicialmente a longitudes de onda cortas, pero más adelante, radiada ya la mayor parte de su energía, emiten sólo a longitudes de onda largas. La disminución en intensidad de las emisiones de radio de Cisne X-3 fue interrumpida por nuevos brotes de actividad de radio, entre ellos un potente destello se produjo aproximadamente un mes después de la explosión inicial. Por lo que parecía, la emisión de partículas energéticas era un proceso activo.

Diez años más tarde, en 1982, se registró otra explosión gigante. Robert Geldzahler y sus colegas, del Naval Research Laboratory, sacaron partido del suceso: con el recién acabado radiointerferómetro de la Gran Disposición (Very Large Array), en Nuevo México, trazaron los primeros mapas, que mostraron que Cisne X-3 era algo más que una fuente puntual. La región de emisión de radio era elíptica y a medida que la explosión progresaba la elipse se hacía gradualmente más elongada. Para Geldzahler y sus colaboradores, la elongación mostraba que Cisne X-3 estaba emitiendo chorros de partículas a velocidades de aproximadamente un tercio de la velocidad de la luz. De los núcleos de algunas galaxias activas emanan chorros de radio, pero sólo se ha observado otra fuente en nuestra galaxia que los emita: el objeto extraño llamado SS 433.

Cabía esperar que una fuente de partículas de alta energía emitiera rayos gamma. Por ello, no sorprendió que los rayos gamma de Cisne X-3 se detectaran durante la explosión de radio de 1972. Usando un instrumento a bordo de un globo, A. M. Galper y sus colegas, del Instituto de Ingeniería Física de Moscú, observaron un fuerte flujo de rayos gamma de baja energía en la dirección de Cisne. La energía de los rayos gamma era de unos 4×10^7 electronvolt. Más tarde, durante el mismo año, el satélite *SAS-2* de la NASA registró radiación gamma con una energía de 10^8 electronvolt. Analizando la señal, Richard C. Lamb, Carl E. Fichtel, Robert C. Hartmann, Donald A. Kniffen y David J. Thompson, del Centro de Vuelos Espaciales Goddard de la NASA, encontraron una periodicidad de 4,8 horas que identificaba sin ambages la señal como procedente de Cisne X-3.



8. IMAGEN DE CISNE X-3 EN EL INFRARROJO PROXIMO (*aspas*). La tomó uno de los autores (Trevor C. Weekes), junto con John C. Geary, del Centro de Astrofísica del Observatorio del Harvard College. Para ello se sirvieron de un instrumento de carga acoplada conectado al Telescopio de Espejo Múltiple, en monte Hopkins (estado de Arizona). La luz visible de Cisne X-3 es absorbida por las nubes de polvo galáctico interpuestas, pero su emisión infrarroja sugiere que brilla intrínsecamente, y bastante, en el dominio del visible.

Pero no resultaba todavía obvio que Cisne X-3 estuviera muy relacionado con el problema del origen de los rayos cósmicos. Los rayos gamma de energías de 10^8 electronvolt se ubican en el extremo inferior del espectro de los rayos cósmicos; podría pensarse en su generación en procesos no asociados con partículas de alta energía. Para ser considerado una fuente significativa de rayos cósmicos, un objeto debe emitir rayos gamma de 10^{12} electronvolt por lo menos.

Los primeros rayos gamma de estas características procedentes de Cisne X-3 fueron observados por Arnold A. Stepanian y sus colegas en el Astrofísico de Crimea. Desde 1970 habían estado buscando fuentes galácticas de rayos gamma de muy alta energía, mediante rastreo de la luz de Cerenkov de cascadas atmosféricas; sus equipos constaban de cuatro espejos reflectores, provenientes de remanentes del ejército, con fototubos en los focos. Al oír las noticias de la explosión de Cisne X-3 en septiembre de 1972, los investigadores de Crimea dirigieron sus espejos hacia Cisne. En 11 noches de observación detectaron un fuerte flujo de rayos gamma de 10^{12} electronvolt. La señal exhibía la periodicidad característica de 4,8 horas, aunque el pico era más agudo que los registrados a otras longitudes de onda; se recibía también en una fase distinta del ciclo.

En esa situación, Cisne X-3 debía haber atraído la atención de los teóricos de los rayos cósmicos, pero no lo hizo. Los resultados del grupo de Crimea cayeron en el olvido, quizás, en parte, porque no se publicaron en el número de *Nature* donde aparecieron la mayoría de trabajos relativos a la explosión. El hecho de que los modelos teóricos de fuentes de rayos X que prevalecían entonces no contemplaran la emisión de rayos gamma de muy alta energía pudo haber contribuido también a su postergamiento. En defensa de los teóricos, hay que reconocer que las observaciones de cascadas atmosféricas habían producido algunas falsas alarmas en el pasado.

Ello no frenó la labor de los astrónomos de Crimea, que continuaron sus observaciones. En cada período de observación entre 1972 y 1979 midieron la señal de Cisne X-3; la amplitud disminuía y cambiaba la forma de la curva de luz (la variación de las emisiones con el tiempo), pero la señal se encon-

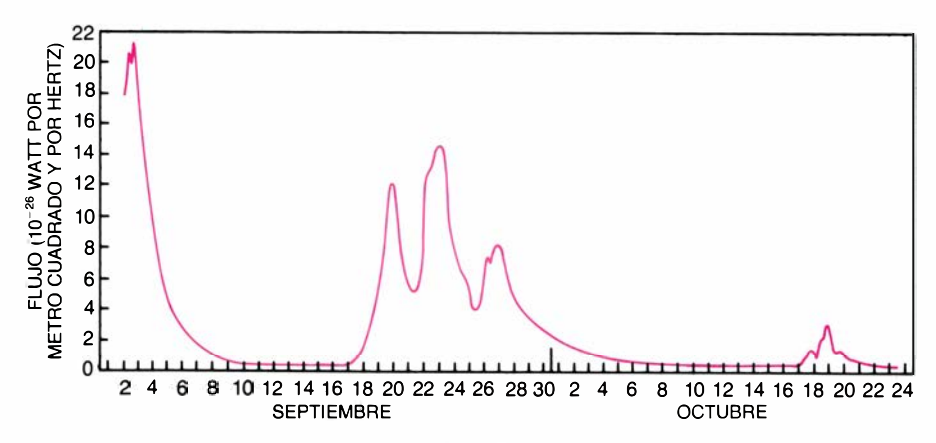
traba siempre allí. Finalmente, en 1980, los resultados de Crimea fueron confirmados por otro grupo independiente: Sean Danaher, David Fegan y Neil A. Porter, del University College de Dublín, en colaboración con uno de nosotros (Weekes) en el Whipple Observatory, detectaron las emisiones de Cisne X-3 con un pequeño telescopio de Cerenkov. En los dos años siguientes, las observaciones de dos grupos más establecieron sin lugar a dudas que Cisne X-3 era una fuente de rayos gamma de 10^{12} electronvolt y, por tanto, de partículas de alta energía.

Desde entonces, los datos obtenidos por redes de detectores de partículas han mostrado que Cisne X-3 emite también rayos gamma de energía ultraalta. Entre 1976 y 1980 los investigadores de la Universidad de Kiel, en Alemania Occidental, mantuvieron en funcionamiento sin interrupción una red de esas. No pretendían tanto buscar una fuente de rayos cósmicos cuanto estudiar las interacciones de alta energía en cascadas atmosféricas generadas por rayos cósmicos en el intervalo de energías de 10^{15} a 10^{16} electronvolt. A fin de determinar la distribución de las partículas secundarias, los investigadores tenían que determinar la dirección de llegada de los rayos cósmicos con una precisión aproximada de un grado de arco. En consecuencia, sus bases de datos sirvieron también para identificar anisotropías direccionales en el flujo de rayos cómiscos.

En 1983, Wilhelm Stamm y Manfred Samorski emprendieron un trabajo laborioso: buscar en los datos de cinco años alguna prueba de emisiones de rayos gamma de Cisne X-3. Descubrieron una nítida abundancia de cascadas atmosféricas originadas aproximadamente en la dirección correcta, pero el margen de error en la determinación de las direcciones de llegada resultaba excesivo para asegurar claramente que Cisne X-3 era la fuente de la señal. La prueba concluyente estaba en los tiempos de llegada de las cascadas. Cuando Stamm y Samorski analizaron los tiempos para las cascadas originadas en la dirección de Cisne X-3, apareció la acostumbrada periodicidad de 4,8 horas.

Poco después de que estos resultados se anunciaran fueron confirmados por un grupo de la Universidad de Leeds. La resolución angular de la red utilizada por los investigadores de Leeds, de 1978 a 1982, no alcanzaba la nitidez obtenida en el experimento de Kiel, pero sí era suficientemente buena para detectar una fuerte señal de Cisne X-3. Además, el grupo de Leeds determinó que Cisne X-3 no presentaba emisiones detectables a energías por encima de los 10^{16} electronvolt.

A pesar de todo, el objeto brilla extraordinariamente a energías ultraaltas. Su flujo observado en el intervalo de energías comprendido entre 10^{15} y 10^{16} electronvolt es de 6×10^{36} erg por segundo, unas 1000 veces la emisión del Sol en todas las longitudes de onda. Añádase a ello que los rayos gamma de energías cercanas a 10^{15} electronvolt son absorbidos en el espacio interestelar; se les supone aniquilados en las colisiones con los fotones de microondas de baja energía que llenan el universo a raíz de la gran explosión. Por tanto, el flujo realmente emitido por Cisne X-3 debe ser aún mayor. Admitiendo que Cisne X-3 está a la distancia mínima de 37.000 años luz, su potencia en rayos gamma de energía ultraalta vendrá a triplicar el flujo observado, unos 2×10^{37} erg por segundo.



9. EXPLOSION DE RADIO de Cisne X-3, en septiembre de 1972. Mostró que este objeto era capaz de acelerar partículas a altas energías. La explosión se descubrió poco antes de que alcanzara el máximo, en cuyo momento el flujo de radio de Cisne X-3 fue mil veces más intenso de lo habitual. Se representa en la figura la explosión a una frecuencia de 8085 megahertz. A frecuencias más bajas (y menores energías) duró más tiempo, un comportamiento característico de la radiación de sincrotrón de electrones relativistas.

Por otro lado, los rayos gamma constituyen sólo una pequeña parte del flujo total de radiación cósmica de Cisne X-3. Los diferentes modelos teóricos del objeto coinciden en que los rayos gamma se emiten probablemente en interacciones de alta energía de núcleos cargados, principalmente protones. En tales interacciones sólo un tres por ciento aproximadamente de la energía liberada se convierte en rayos gamma; el resto se lo llevarían electrones, positrones y neutrinos. Si Cisne X-3 radía 2×10^{37} erg por segundo en rayos gamma de energía ultraalta, debe expulsar por lo menos 30 veces más energía, o sea, 6×10^{38} erg por segundo, en forma de partículas de rayos cósmicos cargadas.

Se trata de una estimación moderada. Por basarse en los rayos gamma observados al llegar a la Tierra, no tiene en cuenta la posibilidad de que Cisne X-3 emita rayos cósmicos en todas direcciones, lo cual es bastante verosímil. Como la potencia galáctica en radiación cósmica está solamente entre 10^{40} y 10^{41} erg por segundo solamente, un pequeño número de fuentes como Cisne X-3 –unas docenas, por ejemplo– explicaría la mayor parte del flujo observado de rayos cósmicos. Quedarían por justificar los rayos cósmicos con energía por encima del límite aparente de Cisne X-3, de 10^{16} electronvolt. Bien podrían venir de otras galaxias.

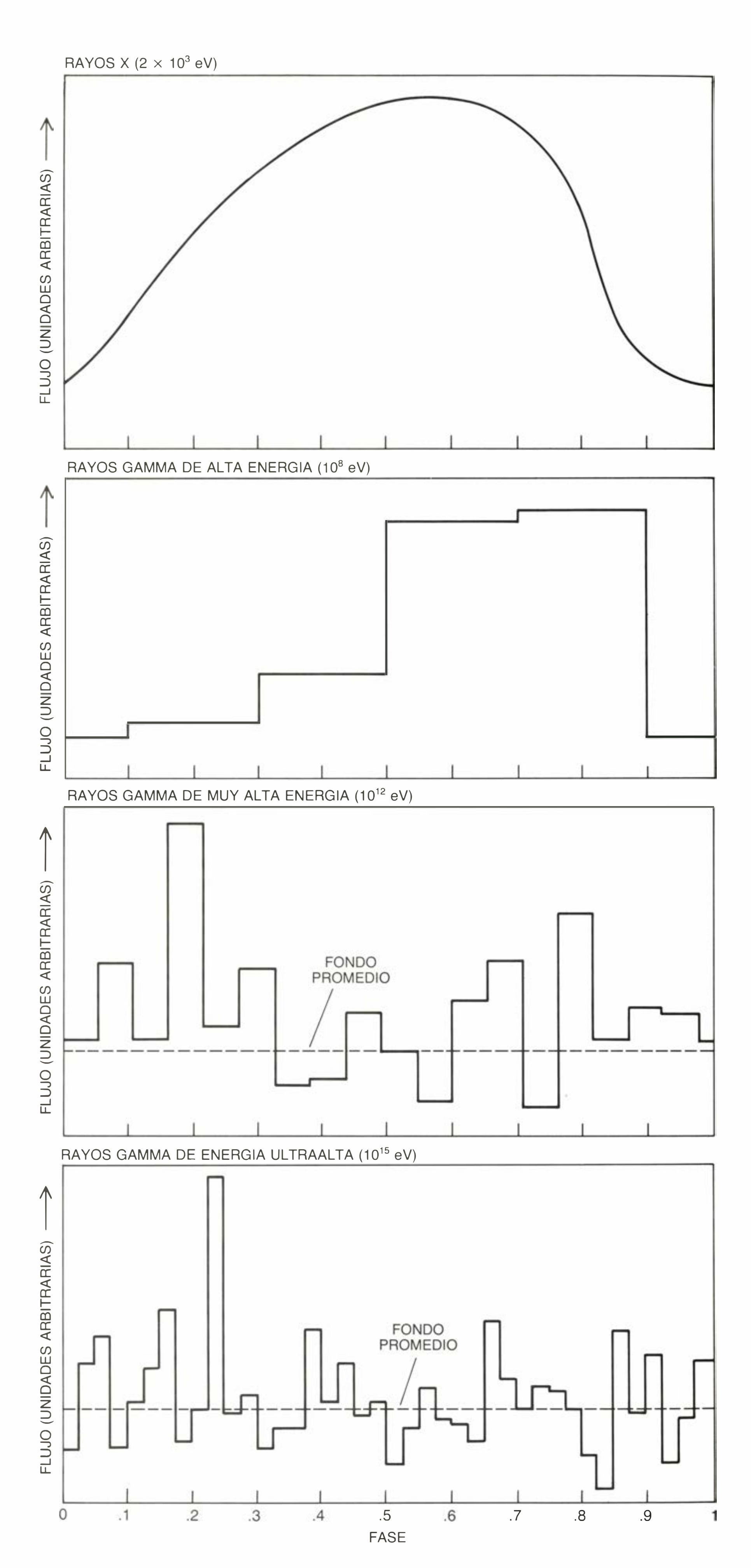
¿A qué se debe que Cisne X-3 sea un acelerador de partículas tan potente? Apoyándose en la periodicidad de sus emisiones, la mayoría de investigadores coinciden en que debe constituir un sistema estelar binario. Las binarias de rayos X no son raras en la galaxia. Se las supone formadas por una estrella densa y colapsada en órbita alrededor de una estrella normal dotada de gran masa. Si las órbitas son suficientemente cercanas, el intenso campo gravitatorio de la estrella densa puede arrancar materia de la superficie de su compañera. A medida que la materia gaseosa cae sobre la estrella densa es acelerada hasta velocidades altísimas; las colisiones entre las moléculas calientan el gas a decenas de millones de grados Celsius; parte de la energía térmica se radía en forma de rayos X.

En el caso de Cisne X-3 la energía necesaria para acelerar partículas hasta las velocidades de los rayos cósmicos debe también venir de la estrella densa, probablemente una estrella de neutrones. Esta posee una masa mayor que el Sol, aunque sólo mida unos diez kilómetros de diámetro; es el núcleo colapsado de una estrella masiva cuyas capas

exteriores han saltado despedidas en una explosión de supernova. El colapso comprime el campo magnético de la estrella y aumenta su intensidad hasta un billón de veces. Crece también la velocidad de rotación de la estrella, pues su momento angular se conserva durante el colapso. En los siguientes milenios esta estrella giratoria puede irse frenando gradualmente.

Según el modelo de Cisne X-3 propuesto por Thomas Vestrand, de la Universidad de New Hampshire, y Davis Eichler, de la Universidad de Maryland en College Park, la estrella de neutrones se hallaría todavía girando rápidamente: sería un púlsar. La energía impartida a los rayos cósmicos derivaría de la energía cinética rotacional perdida por el púlsar en su frenado. Los púlsares emiten radiación de sincrotrón y podrán, en consecuencia, acelerar electrones, a energías relativistas por lo menos. Aunque se ignoran los pormenores del proceso, el campo magnético giratorio induciría un campo eléctrico que podría entonces acelerar partículas cargadas –protones y electrones– de la superficie del púlsar.

De acuerdo con el modelo propuesto por Vestrand y Eichler, el púlsar de Cisne X-3 arroja protones de alta energía en todas direcciones. La mayoría de los cuales se dispersarían por el espacio, pero algunos penetrarían en la capa gaseosa exterior de la estrella acompañante. Cuando un protón choca con un núcleo de gas, produce tres clases de piones, una de las cuales se desintegra en dos rayos gamma. Los rayos gamma siguen una trayectoria próxima a la inicial del protón, y el modelo predice que formarán un haz dirigido hacia la Tierra justamente en dos puntos de la órbita del púlsar en torno a la estrella acompañante. Esto explicaría la curva de luz con dos picos gemelos registrada en algunas observaciones de Cisne X-3 a energías de 10^{12} electronvolt. A. Michael Hillas, de Leeds, ha demostrado, además, que el modelo del púlsar puede explicar la intensidad observada de los rayos gamma emitidos



10. PERIODICIDAD de las emisiones de rayos gamma de Cisne X-3. Se ajusta a la periodicidad de 4,8 horas observada a longitudes de onda de rayos X. Dicha regularidad se atribuye a la órbita de la estrella binaria. Las curvas de flujo de los rayos gamma son promedios obtenidos superponiendo emisiones de intervalos de tiempo sucesivos de 4,8 horas. Si la señal variara al azar, la superposición daría una curva plana. Los picos de la superposición implican que los picos de la señal son recurrentes, con la misma fase de los sucesivos intervalos de tiempo; así que la señal es periódica. La diferencia de fase entre la emisión de 10^8 electronvolt y la emisión a energías más altas sugiere que las señales vienen de regiones distintas del sistema binario.

por Cisne X-3 en todas las energías, desde 10^{12} hasta 10^{16} electronvolt, en el supuesto de que el púlsar acelere protones a 10^{17} electronvolt.

Esta energía centuplica la considerada habitualmente límite superior de la capacidad de aceleración de un púlsar. El modelo de Vestrand-Eichler requiere que los púlsares se muestren mucho más eficaces, a la hora de acelerar partículas, de lo que se venía sosteniendo. También supone que el púlsar de Cisne X-3 gira con vertiginosa rapidez, entre 100 y 1000 veces por segundo. Hasta ahora sólo se han observado dos púlsares con un ritmo de rotación de esta magnitud. Obviamente el descubrimiento de un púlsar rápido en Cisne X-3 daría mucha credibilidad al modelo del púlsar.

Un segundo modelo, propuesto por Ganesar Chanmugam, de la Universidad estatal de Louisiana, y Kenneth Brecher, de la Universidad de Boston, no exige que la estrella de neutrones rote a esas velocidades. En este caso, la energía que se convierte en rayos cósmicos proviene del proceso que es también responsable de la emisión de rayos X: la acreción de materia de la compañera masiva sobre la estrella de neutrones. Cuando el disco de acreción gira en torno a la estrella de neutrones, el campo magnético de ésta induce un campo eléctrico en el plano del disco. Los protones reciben una aceleración hacia afuera, en dirección radial, a lo largo de las líneas del campo eléctrico. Algunos de los protones penetran en la corriente de acreción y allí chocan contra los núcleos del gas generando rayos gamma. Este modelo presenta también sus dificultades. Verbigracia: se exige un campo eléctrico intenso, y no está claro si un campo así puede mantenerse en un disco gaseoso.

El modelo de Chanmugom-Brecher explicaría por qué las emisiones de rayos gamma de alta energía parecen haberse detectado en otros dos sistemas binarios de rayos X (Vela X-1 y Hércules X-1) que no contienen púlsares rápidos. Sin embargo, la decisión de escoger entre los dos modelos de producción de rayos cósmicos, así como llegar a una comprensión final de Cisne X-3, quizá tarden en producirse. Queda mucho por averiguar y no hay duda de que habrá sorpresas.

A modo de ejemplo, la radiación cósmica de Cisne X-3 podría enseñar a los investigadores cosas nuevas acerca de la física de partículas. Aunque el exceso de cascadas atmosféricas que llegan en la dirección de Cisne X-3 puede atribuirse a la radiación gamma de alta energía, nadie ha identificado inequívocamente a los rayos gamma como causantes del exceso; se llega a esta conclusión por eliminación de los neutrones y los neutrinos, los otros candidatos de que se tiene noticia. Pero, ¿se conocen todos los demás candidatos? ¿Se conoce la física de las interacciones a 10^{15} electronvolt, energía esta mucho mayor que la alcanzada en los aceleradores construidos por el hombre? ¿Podría el flujo de la radiación cósmica de Cisne X-3 contener un nuevo tipo de partícula?

Informes recientes de dos experimentos de desintegración del protón, uno de la mina de Soudan, en Minnesota, y otro en el túnel del Montblanc, sugieren por lo menos algo nuevo. Ambos experimentos tienen por misión detectar los productos de la desintegración del protón, que pueden dar lugar a muones (partículas cargadas negativamente, de vida corta). En consecuencia, los aparatos registran también la dirección de llegada de los muones generados en cascadas atmosféricas de rayos cósmicos. A partir de estos datos se pueden inferir las direcciones de llegada de los propios rayos cósmicos.

Tras analizar los datos registrados a lo largo de un año, el grupo de Soudan ha informado de un exceso de rayos cósmicos en la dirección de Cisne X-3. La señal está marcada por la característica periodicidad de 4,8 horas. Una periodicidad similar se observa en los datos del experimento del Montblanc. Sin embargo, se espera que las cascadas atmosféricas provocadas por los rayos gamma contengan muy pocos muones y no produzcan una señal medible en estos detectores. Los resultados sugieren así que alguna partícula desconocida emitida por Cisne X-3 puede ser la causa de las cascadas atmosféricas. O bien, adoptando una postura más conservadora, quizá no comprendamos las interacciones de los rayos gamma de alta energía y, después de todo, puedan producir muchos muones.

Tales cuestiones abiertas no deben oscurecer nuestro punto central: por primera vez se ha identificado una fuente significativa de rayos cósmicos dotados de altísima energía. Tras décadas de lucubraciones sin fundamento apenas, los astrofísicos tienen ahora un prototipo de acelerador de rayos cósmicos, a pesar de que no comprendan del todo su funcionamiento. En los próximos años, telescopios de rayos gamma más sensibles detectarán, a lo mejor, nuevas fuentes de rayos cósmicos que sean versiones más débiles de Cisne X-3.

La señal del calcio

El ion de calcio controla gran variedad de procesos: desde la contracción muscular hasta la división celular. Un equipo de proteínas, especializadas en unirse al calcio, regulan su concentración intracelular y controlan sus efectos

Ernesto Carafoli y John T. Penniston

Un detonador común desencadena procesos biológicos tan diversos como la contracción muscular y la secreción hormonal. El desencadenante es un pequeñísimo flujo de iones de calcio, elemento que se cuenta entre los segundos mensajeros de los seres vivos: transporta los mensajes químicos y eléctricos desde la membrana plasmática, donde alcanzan la célula, hasta la maquinaria bioquímica intracelular. La consecución de un control eficaz de los procesos celulares requiere, a su vez, la propia regulación del calcio intracelular. Con ese fin las células han desarrollado un elaborado sistema de proteínas que, interactuando con el calcio, gobiernan la transmisión y recepción del mensaje intracelular. El conocimiento adecuado de esos intrincados procesos permitiría un mejor control clínico del calcio intracelular, posibilidad que sin duda repercutiría considerablemente en el tratamiento de diversas enfermedades.

Son varios, además del calcio, los iones contenidos en los fluidos biológicos. Entre otros, magnesio, con doble carga, y sodio, potasio y cloro, de carga simple. ¿Por qué, en el curso de la evolución, se seleccionó el calcio como segundo mensajero? Para que una sustancia actúe de mensajero intracelular debe unírsele con gran afinidad y especificidad una proteína efectora (diana), habitualmente una enzima. La unión del mensajero a la proteína altera la conformación de la molécula enzimática, y con ello su estado de actividad. La concentración del mensajero debe sufrir amplias oscilaciones para que se altere el número de moléculas diana afectadas. Puede requerirse que la concentración intracelular del mensajero aumente diez veces para que una enzima intracelular pase del estado de no activación al de activación.

Ni siquiera un ion de la simplicidad del calcio se crea o se destruye fácilmente. Antes bien, para que actúe de mensajero deben regular su concentración en el citosol (fluido viscoso donde están suspendidos los orgánulos intracelulares) componentes celulares que, alternativamente, lo secuestren, eliminándolo del citosol, y lo viertan a él de nuevo. Tales componentes deben presentar cierta complejidad estructural que les permita distinguir el ion mensajero de otros iones intracelulares y, además, secuestrarlo eficazmente. Toda función celular que, para su realización, exija la participación de sustancias estructuralmente complejas suele encomendársele a una proteína.

De entre los iones frecuentes en los medios biológicos, el calcio es el mejor dotado para unirse y enlazarse específicamente. Los radios iónicos del cloruro y el potasio, relativamente grandes, les impiden encajar en los compactos centros de unión de las proteínas. El radio iónico del sodio es mucho más pequeño, parecido al del calcio, pero, dado que sólo posee una carga, su unión a las proteínas es muy débil. Los iones poliatómicos comunes en los medios biológicos, así fosfato o bicarbonato, son todavía mucho mayores, menos capaces, por tanto, de formar complejos estables.

Tras esos procesos de eliminación, los únicos candidatos restantes a segundos mensajeros son los iones de calcio y magnesio. Ambos poseen un radio iónico pequeño, presentan dos cargas y son capaces de formar uniones estables con otros compuestos. ¿Cómo explicar la preferencia evolutiva por el calcio? Robert J. P. Williams, de la Universidad de Oxford, ha estudiado las características químicas de la unión del calcio y el magnesio. Cuando establecen complejos con las proteínas, ambos iones se unen a seis dadores de electrones, usualmente átomos de oxígeno, adoptando una disposición octaédrica. Un dador ocupa cada vértice del octaedro, y los enlaces adyacentes se disponen formando ángulos rectos. El ion calcio, que es mayor que el de magnesio y posee una estructura electrónica más compleja, puede establecer uniones con un total de siete u ocho dadores de electrones.

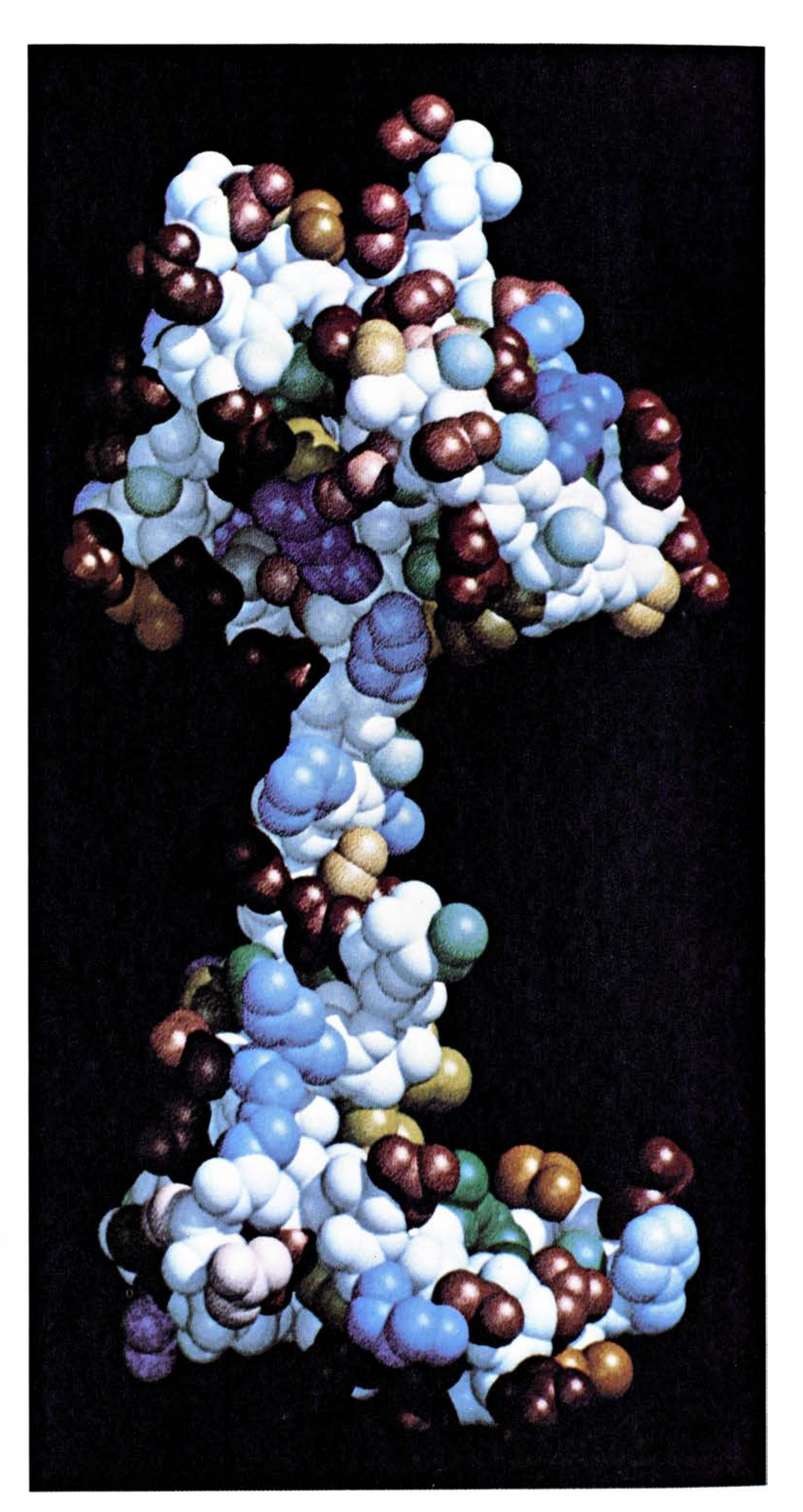
El magnesio, debido a su pequeño tamaño, tiende a tirar de los oxígenos de la proteína con la que forma complejo, originando una configuración regular y rígida. Las proteínas, sin embargo, no suelen ser lo suficientemente flexibles para crear una cavidad regular adecuadamente compacta que se ajuste exactamente a las pequeñas dimensiones del ion magnesio. En vez de establecer las seis uniones con la proteína, el magnesio también se enlaza con moléculas de agua. Tal sustitución debilita enormemente la fuerza del engarce, puesto que para liberar al magnesio se requerirá romper un menor número de enlaces entre el ion y la proteína. La unión del calcio, por el contrario, requiere un cambio de conformación de la proteína menos drástico, lo que permite al ion establecer los seis enlaces con la proteína.

La unión del calcio a las proteínas no sólo resulta más fuerte que la del magnesio, sino también más específica. Gracias a su mayor radio iónico y a la capacidad de establecer un número variable de enlaces, el calcio puede encajarse en centros de unión de conformaciones irregulares. De ahí que una proteína pueda unirse al calcio de forma compacta, excluyendo al magnesio, a pesar de que la concentración citosólica de éste es mil veces mayor que la del calcio. Por el contrario, una proteína que forme complejo con el magnesio acogerá al calcio por igual, si no mejor. Tanto la hermeticidad como la especificidad de la unión resultan esenciales para que una sustancia actúe como segundo mensajero; sólo el calcio reúne ambos requisitos.

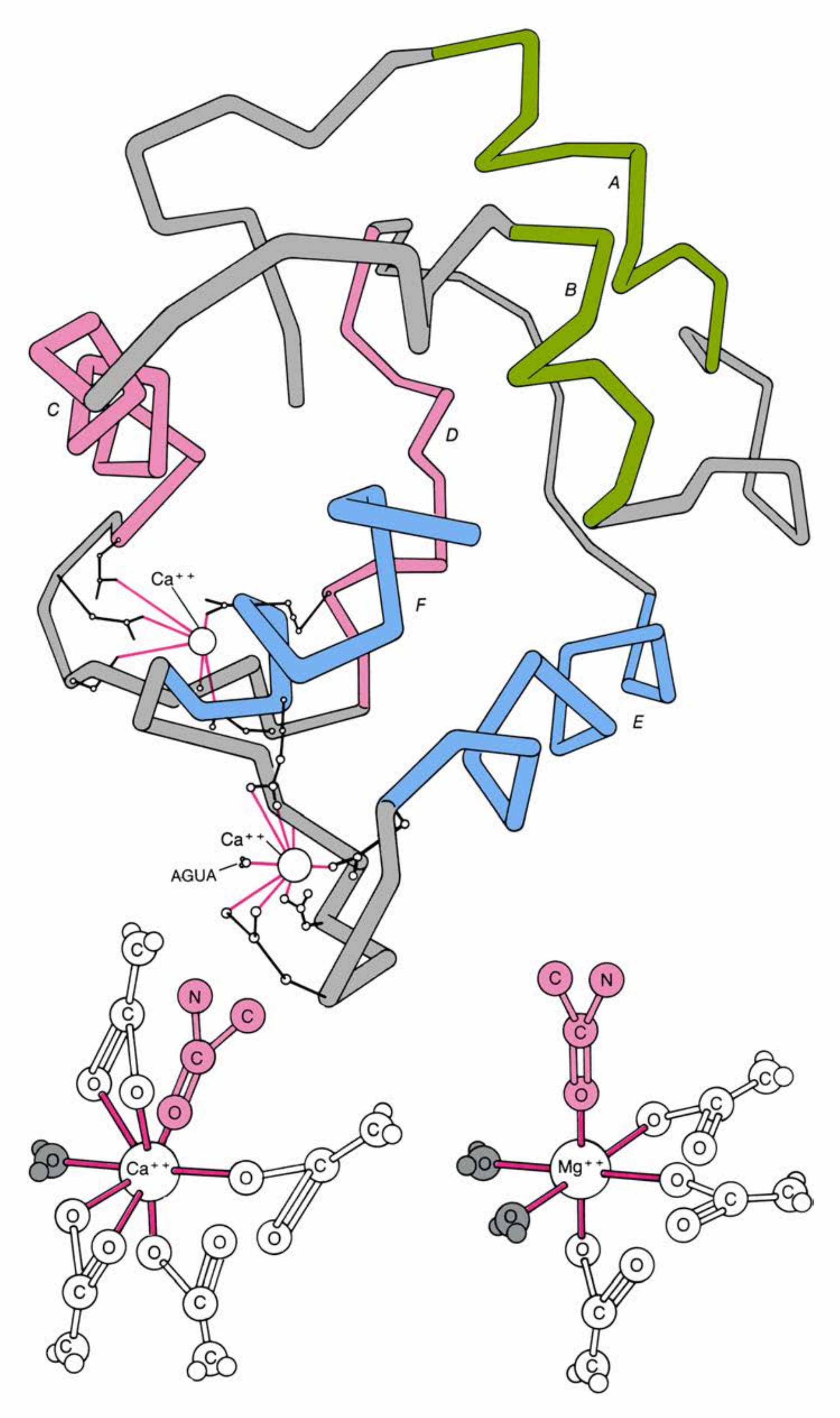
Se distinguen dos clases de proteínas capaces de formar complejos con el cal-

cio: proteínas incorporadas en las membranas de las células, que gobiernan el paso de calcio al citoplasma e interior de orgánulos intracelulares (y en sentido inverso), y proteínas solubles en el citoplasma e interior de los orgánulos. Esta segunda clase de proteínas desempeña cierto papel en el control de la concentración citosólica de calcio ionizado, aun cuando la cantidad de calcio que puede formar complejo con una proteína soluble está limitada por el número de moléculas proteicas. Los polipéptidos inmersos en las membranas celulares pueden regular con mayor eficacia la concentración citosólica de calcio, ya que actúan de transportadores. Así, una sola molécula de proteína puede captar el ion a un lado de la membrana, transferirlo al otro lado, liberarlo y repetir el proceso de forma indefinida. La principal función de las proteínas solubles en el citosol y en el interior de los orgánulos celulares es, por el contrario, la de mediar en los efectos intracelulares de la señal.

En algunos casos, una proteína citosólica que forme complejo con el calcio es, por sí sola, responsable de los efectos de la señal del calcio. Normalmente, la proteína captadora del calcio actúa como intermediaria, activando otras enzimas. En 1967, Setsuro Ebashi y sus colaboradores, de la Universidad de Tokio, identificaron una proteína de ese tipo. Describieron un polipéptido que permitía al calcio desencadenar la contracción del músculo esquelético, denominándolo troponina. Estudios posteriores revelaron que la troponina es una molécula compleja, compuesta de tres subunidades proteínicas, y sólo una de ellas, la troponina *C*, es la receptora de calcio. Cuando la célula muscular se halla en reposo, parte de la molécula de troponina actúa de inhibidor, impidiendo que la miosina, la proteína contráctil del músculo, se una a la actina, la proteína filamentosa.



1. PROTEINA QUE SE UNE AL CALCIO. Como todas las que muestran esa capacidad, resulta esencial para que el calcio actúe de mensajero intracelular. Algunas de las proteínas que interactúan con el calcio controlan su concentración intracelular, y así generan señales; otras captan la señal. En la figura se muestra la troponina *C*, proteína de los músculos esquelético y cardiaco. Cuando el calcio se une a la troponina *C* se produce un cambio conformacional de la proteína, que desencadena la serie de procesos fisiológicos que conducen a la contracción muscular. En esta imagen de ordenador, las esferas se corresponden con átomos individuales; las regiones coloreadas representan cadenas laterales de diversas características químicas. La imagen es obra de Osnat Herzberg y Michael N. G. James, de la Universidad de Alberta, que se basaron en datos obtenidos de análisis de la proteína por rayos X.



2. PARVALBUMINA. Presente en peces, reptiles y anfibios, es la típica proteína que se une al calcio con gran especificidad y firmeza. La molécula (*arriba*) posee seis regiones helicoidales (*color*) entremezcladas con tres dominios de unión al calcio. Un par de hélices y el dominio que flanquean constituyen un "asa". El asa que contiene las hélices *A* y *B* no es funcional. Los iones calcio están unidos a las otras dos asas, las definidas por las hélices *C* y *D* y *E* y *F*. En una esquina de la figura (*abajo, a la izquierda*) se muestra en detalle la unión del calcio al asa *EF*. El ion se une a seis átomos de oxígeno de las cadenas laterales, y a otro del eje central de la proteína (*color*). Un octavo enlace se establece con el oxígeno del agua (*gris*). La longitud de las uniones y los ángulos entre enlaces adyacentes varían. Un intento hipotético de encajar al ion magnesio, el principal competidor del calcio por los centros de unión de la proteína, en una de las asas (*abajo, a la derecha*) muestra por qué el asa presenta una afinidad muy superior por el calcio. La estructura electrónica del ion magnesio y su pequeño tamaño molecular exigen que los seis enlaces formen un octahedro compacto. Dado que la parvalbúmina no es una molécula lo suficientemente flexible, el ion debe sustituir dos de sus uniones con la proteína por enlaces con moléculas de agua, debilitándose la unión. (Los dibujos son de George V. Kelvin.)

Cuando el calcio se une a la subunidad *C* de la troponina, ésta sufre un cambio conformacional, de tal manera que la subunidad inhibidora se desplaza hacia un lado, permitiendo la unión de actina y miosina, y en consecuencia la contracción del músculo.

El estudio de una segunda proteína citoplasmática, la parvalbúmina, presente en grandes cantidades en los músculos esqueléticos de peces, reptiles y anfibios, permitió esclarecer cómo se une el calcio a las proteínas. Aunque no se conoce con precisión la función de la parvalbúmina, Robert H. Kretsinger y colaboradores, de la Universidad de Virginia, determinaron su estructura tridimensional valiéndose de técnicas cristalográficas de rayos X. La molécula de parvalbúmina posee tres regiones muy similares, cada una de las cuales forma un bucle de 10 a 12 residuos de aminoácidos, a los que se une el calcio. Cada bucle está flanqueado por dos secuencias helicoidales de una longitud aproximada de diez aminoácidos. El interior de cada bucle receptor de calcio posee de seis a ocho átomos de oxígeno, de las cadenas laterales y del eje peptídico de la proteína, que donan electrones para unir el ion. Uno de los bucles es inactivo, porque dos de los aminoácidos necesarios para la unión con calcio se han remplazado por otros.

Hasta que recientes estudios cristalográficos revelaron la estructura de otras proteínas acomplejantes de calcio, sólo se conocía su secuencia aminoacídica. La mera comparación con la secuencia aminoacídica de la parvalbúmina, sin embargo, permitía anticipar los resultados de los estudios cristalográficos. En esas comparaciones, la región proteínica formada por el bucle receptor del calcio y las hélices *E* y *F* que lo flanquean (denominadas, en conjunto, "asa *EF*") sirve de modelo para la búsqueda de los centros de unión de calcio en otras proteínas.

Tales estudios han demostrado que las demás proteínas acomplejantes de calcio suelen parecerse a la parvalbúmina; en particular, poseen más de un dominio receptor de calcio. Es probable que la forma de la proteína cambie atendiendo al número de centros de unión ocupados. Cada una de las formas resultantes podría interactuar con una proteína efectora ("diana") distinta, o bien ejercer distintos efectos sobre una misma proteína diana.

Se cree que la calmodulina, la tercera de las más importantes proteínas receptoras de calcio, funciona de esa

manera. La calmodulina posee cuatro centros de unión, y recientes estudios cristalográficos han mostrado que todos ellos se parecen al asa *EF* de la parvalbúmina. Las propiedades de la molécula de calmodulina con sólo uno o dos lugares ocupados pueden diferir de las que presenta la molécula cuando están ocupados por calcio sus cuatro centros. Tal versatilidad de las propiedades de la calmodulina explicaría el vasto número de proteínas con las que interactúa. La calmodulina parece ser uno de los más comunes traductores del mensaje del calcio, presente en todos los tipos de células de mamífero.

La calmodulina actúa sobre enzimas que catalizan la formación o ruptura de enlaces entre fósforo y otros átomos frecuentes en la bioquímica celular. Entre tales enzimas se cuentan las quinasas, que fosforilan (añaden grupos fosfato a) otras proteínas. A menudo, la fosforilación actúa de conmutador que activa o inactiva una enzima. La calmodulina, por tanto, regula numerosas funciones celulares. Investigaciones recientes sugieren que la calmodulina podría estar implicada en la liberación de hormonas de glándulas endocrinas. Parece también controlar la forma y división celulares, actuando probablemente sobre los microtúbulos, delgados túbulos de tubulina, una proteína, que constituyen el entramado intracelular. La calmodulina puede, por ejemplo, desencadenar la ruptura de microtúbulos ante la presencia de otras proteínas. Dado que las células cancerosas difieren de las normales fundamentalmente en su forma y velocidad de división celular, la calmodulina activada por calcio bien pudiera desempeñar algún papel en la bioquímica de dicha enfermedad.

La función reguladora que realiza la calmodulina en músculo esquelético se conoce bastante mejor. Un aumento de la concentración intracelular de calcio ionizado aumenta instantáneamente una respuesta en la que interviene la troponina: la contracción muscular. También provoca, sin embargo, un cambio metabólico a largo plazo, mediado por la calmodulina. La calmodulina activada por calcio activa a su vez una proteína quinasa, que fosforila una segunda enzima. Tras su activación, la segunda enzima cataliza la degradación del glucógeno (forma de almacenamiento de la glucosa) en glucosa. Del metabolismo de la glucosa se obtiene la energía necesaria para el trabajo muscular.

La unión del calcio a proteínas resulta de la mayor importancia tanto para la generación de la señal intracelular del calcio como para su recepción. La concentración de calcio ionizado, en el citoplasma e interior de orgánulos intracelulares, está controlada principalmente por proteínas que se localizan en la membrana plasmática de la célula y en las membranas de dos tipos de orgánulos celulares: retículo endoplasmático y mitocondria. Las oscilaciones de la concentración citosólica de calcio constituyen el mensaje del calcio, y las proteínas regulan dichas oscilaciones captándolo y vertiéndolo o retirándolo del citoplasma. Las proteínas también permiten mantener la concentración citosólica de calcio ionizado en niveles infinitesimales. En una célula típica de mamífero, la concentración citosólica de calcio es del orden de 0,1 micromolar, esto es, de cuatro millonésimas de gramo por litro, alrededor de 10.000 veces inferior a la del plasma sanguíneo. (La molaridad de una solución equivale al número de gramos de soluto por litro, dividido por el peso atómico, o molecular, del soluto.)

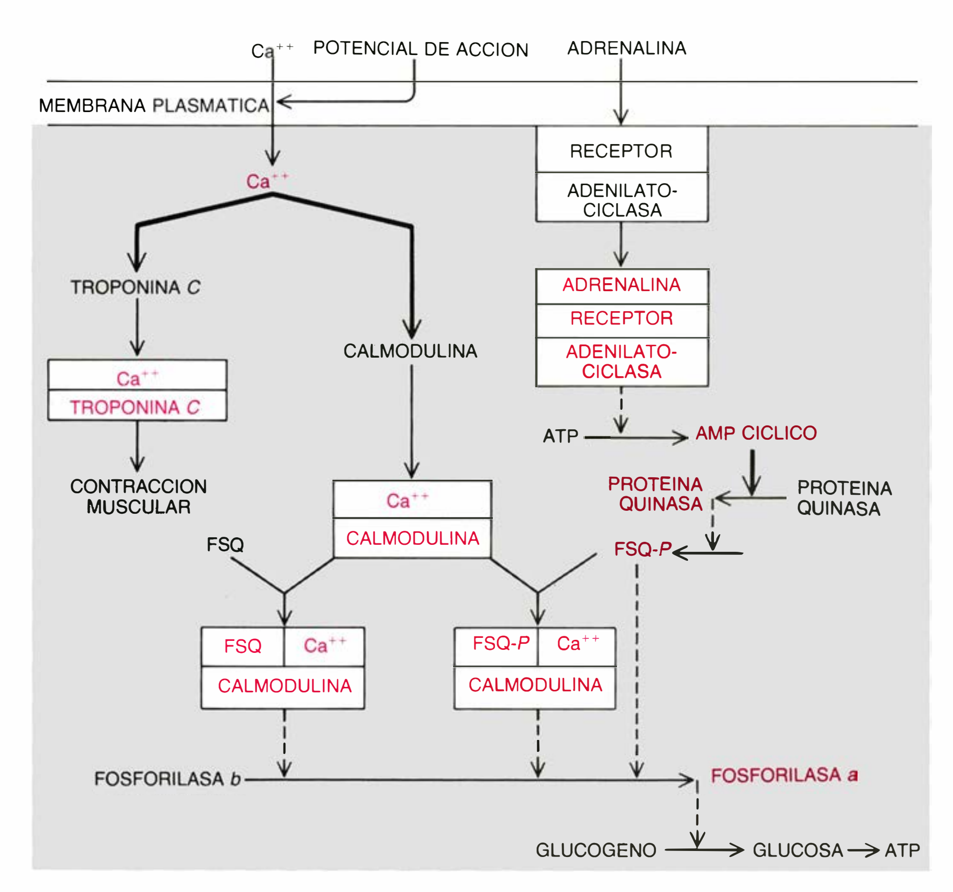
En 1976, Annemarie Weber, de la Universidad de Pennsylvania, destacó que el mantenimiento de una baja concentración intracelular de calcio ionizado es condición necesaria para el funcionamiento del metabolismo del fosfato, característico de los animales superiores. Al adenosín trifosfato (ATP) le corresponde el suministro energético de la mayoría de los procesos celulares. Su ruptura libera fosfato inorgánico. Si la concentración intracelular de calcio ionizado fuera alta, el fosfato y el calcio se combinarían en un precipitado de cristales de hidroxiapatita, sustancia pétrea que forma parte del cemento del hueso. En última instancia, la calcificación mataría la célula.

Una baja concentración citosólica de calcio ionizado permite disponer de un mensajero intracelular de bajo coste energético. El movimiento de calcio a través de las membranas requiere energía que, normalmente, se obtiene a partir del ATP. Si el nivel intracelular de calcio fuera alto, debería transportarse gran cantidad de iones calcio al interior del citoplasma para que su concentración intracelular se elevara 10 veces, incremento usualmente requerido para activar una enzima; habría luego que expulsar de la célula ese exceso de calcio. La concentración citosólica habitual de calcio, muy baja, permite que sólo deban transportarse a través de las membranas unos pocos iones calcio. En consecuencia, el coste energético que conlleva la regulación de una enzima es relativamente bajo. Por el contrario, el coste energético de la regulación llevada a cabo por otros mensajeros intracelulares, tales como el monofosfato cíclico de adenosina (AMPc), es alto: debe sintetizarse y destruirse AMPc cada vez que transmite un mensaje, y ambos procesos metabólicos requieren un considerable aporte energético.

Las proteínas que controlan el calcio intracelular, localizadas en las membranas, constituyen al menos siete sistemas de transporte diferentes, que abarcan cuatro mecanismos bioquímicos distintos. Se localizan tanto en la membrana plasmática que delimita la célula como en las membranas de orgánulos intracelulares: retículo endoplasmático (presente en la mayoría de las células), retículo sarcoplasmático (típico de las células musculares) y mitocondrias. Los transportadores también difieren en su función: los hay que movilizan calcio hacia el citosol en respuesta a un estímulo externo, generándose entonces la señal del calcio; otros lo expulsan del citosol, manteniendo una baja concentración citosólica de calcio ionizado, lo que confiere eficacia al sistema de mensajes.

Entre los transportadores que expulsan calcio del citoplasma se advierte otra diferencia. Algunos sólo unen y translocan el ion cuando la concentración citosólica del calcio ha alcanzado un valor del orden de uno micromolar, muy por encima de su nivel de reposo (0,1 micromolar). Esos transportadores pueden movilizar grandes cantidades de iones calcio. Probablemente les corresponda la función de eliminar el ion de células dañadas o estimuladas. Otros transportadores son capaces de unirse al calcio cuando la concentración del ion es muy inferior, pero, en contraposición, poseen baja capacidad de transporte. Estos últimos transportadores, de alta afinidad y baja capacidad, son los que realizan un ajuste fino de la concentración citosólica de calcio en células no estimuladas.

La membrana plasmática de la célula dispone de dos transportadores para expulsar el calcio. El sistema de alta afinidad es una ATPasa: enzima que rompe el ATP para obtener energía, que permite al sistema translocar calcio fuera de la célula en contra del pronunciado gradiente de concentración del ion que se registra a través de la membrana plasmática. El transporta-



3. EN EL MUSCULO ESQUELETICO EL CALCIO desempeña una doble función. El potencial de acción, desencadenado en la membrana plasmática por un impulso nervioso, hace que el calcio se difunda hacia el interior de la célula (*flechas gruesas*). El ion interactúa entonces con dos proteínas receptoras, la troponina *C* y la calmodulina. La unión con el calcio insta a las proteínas a pasar del estado inactivo (*negro*) al activo (*color*). La troponina *C* provoca la contracción muscular. La calmodulina se une con una fosforilasa-sintetasa-quinasa (FSQ), inactiva, activándola parcialmente (*color claro*). La FSQ, a su vez, cataliza la conversión de fosforilasa *b* en su forma activa, fosforilasa *a*. (Las flechas a trazos indican procesos catalíticos.) La fosforilasa *a* cataliza la ruptura del glucógeno en glucosa, que se metaboliza para generar ATP, que aporta la energía necesaria para la contracción muscular. Los iones de calcio actúan sinérgicamente con el AMPc, otro mensajero intracelular. Cuando una hormona, como la adrenalina, se une a su receptor de membrana, la adenilato ciclasa cataliza la síntesis de AMPc. Este se difunde hacia el interior de la célula y desencadena una serie de etapas enzimáticas que conducen a la fosforilación de la FSQ. La FSQ fosforilada (FSQ-*P*) sólo es parcialmente activa, activándose por completo cuando se une al complejo de calcio y calmodulina.

dor engloba, en una sola molécula, toda la maquinaria necesaria para acarrear calcio: una porción escinde el ATP y, otra, une y transloca el ion. La bomba la descubrió en los glóbulos rojos H. J. Schatzmann, de la Universidad de Berna. Desde entonces, estudios realizados por nosotros y otros investigadores han demostrado que dicha bomba existe virtualmente en todos los tipos de tejidos de mamífero.

Sirviéndonos de uno de los mecanismos de regulación de la ATPasa bombeadora de calcio logramos, en colaboración con Verena Niggli, del Instituto Federal suizo de Tecnología de Zurich, purificar la proteína de la membrana de los glóbulos rojos. Cuando el calcio intracelular se une a la calmodulina, ésta se enlaza a la bomba de calcio y la activa. A su vez, la bomba reduce la concentración intracelular del ion. Pese a que la membrana del eritrocito sólo contiene 5000 bombas de calcio, lo que constituye la milésima parte del total de proteína presente en la membrana, la especificidad y estabilidad de la unión de la calmodulina a la bomba permite aislar ese complejo del resto proteico de la membrana del glóbulo rojo.

Unimos calmodulina a un soporte sólido y la expusimos a un extracto de membranas de glóbulos rojos humanos. La ATPasa se adhirió a la calmodulina inmovilizada, tornándose temporalmente insoluble. A continuación lavamos el complejo proteico insoluble con una solución de calcio, lo que permitió eliminar otras proteínas. Un lavado con otra solución, en este caso carente de calcio, liberó el calcio de la calmodulina, con lo cual quedó en libertad la ATPasa a ella unida. Posteriormente incorporamos la bomba purificada en membranas artificiales y comprobamos que todavía era capaz de escindir el ATP y bombear calcio.

Niggli examinó las propiedades de la bomba purificada y observó que, al tiempo que expulsaba calcio de la célula, introducía protones (iones hidrógeno) en ella, pròbablemente en una proporción de dos protones por cada ion calcio translocado. Dado que el calcio posee dos cargas positivas, el intercambio de dos protones por un calcio hace que la bomba sea eléctricamente neutra. Por tanto, no le afecta el gradiente eléctrico que se registra a través de la membrana plasmática de muchas células.

Nuestros estudios con la ATPasa purificada revelaron que, además de la calmodulina, también pueden activarla otras sustancias. Una de ellas es un lípido acídico de la membrana plasmática, el denominado fosfatidil-inositol bifosfato. En algunas células se ha observado que, tras estimulación por ciertas hormonas, la presencia de fosfatidil-inositol bifosfato se reduce drásticamente. Puesto que esas mismas hormonas provocan también la entrada de calcio en las células, una disminución de la cantidad de lípido estimulante de la bomba generaría en las células estimuladas hormonalmente una concentración citosólica de calcio muy superior a la normal. Por su alto contenido en calcio, esas células se estimularían por segunda vez mucho más fácilmente. Las células recordarían así su pasado estimulatorio inmediato, hecho de gran importancia para ciertas células nerviosas.

También parece probable que la escisión de la ATPasa por parte de enzimas proteolíticas (enzimas que degradan proteínas) constituya otro mecanismo regulador celular. La bomba de calcio de la membrana de los glóbulos rojos es una de las mayores proteínas de cadena única que se conoce. Con un peso molecular de 138.000, una reducción de hasta 50.000 no bloquea su actividad. De hecho, la eliminación del fragmento proteico suplementario activa la bomba purificada igual que lo hace la calmodulina o el lípido acídico. Ese segmento polipeptídico aparentemente supérfluo probablemente desempeñe un papel regulador, o intervenga en otras funciones de la molécula, por el momento desconocidas.

En la membrana plasmática de músculo cardiaco se ha demostrado la presencia de un cuarto sistema regulador de la ATPasa bombeadora de calcio. En este caso se trata de la fosforilación de la bomba a cargo de una proteína quinasa, activada a su vez por AMPc. Si bien se desconoce cuál sea, en las células vivas, la importancia relativa de los cuatro mecanismos de activación de la bomba de calcio de la membrana

plasmática, su número da testimonio de la importancia de dicha bomba en la fisiología celular.

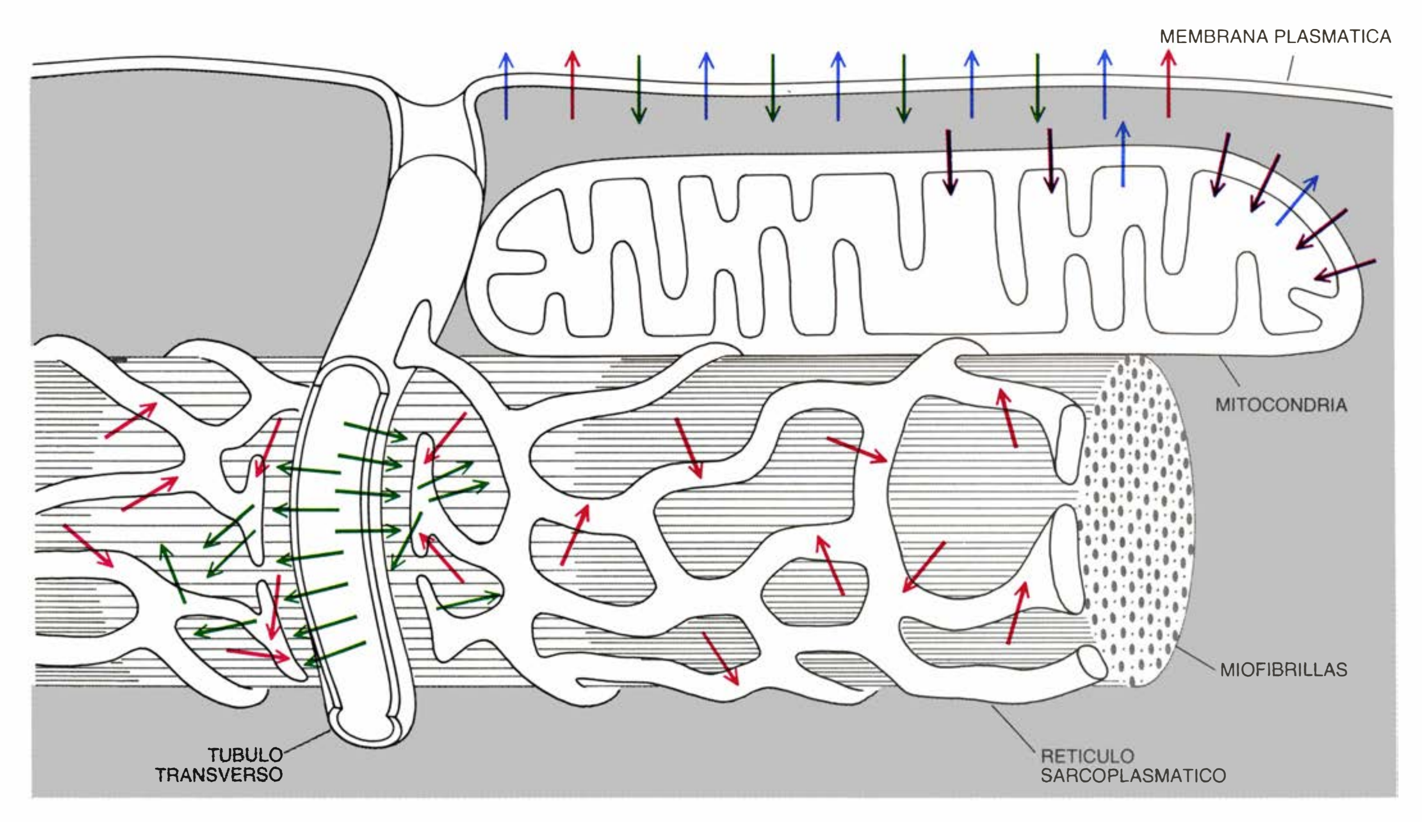
La ATPasa de la membrana plasmática, un transportador de alta afinidad, según se ha visto, responde a minúsculos aumentos de la concentración citosólica de calcio. Oscilaciones más drásticas de la concentración de calcio citosólico activan el otro sistema de que dispone la membrana plasmática para exportar calcio celular, el denominado sistema de intercambio sodio-calcio. Esta proteína intercambiadora abunda particularmente en células excitables: células nerviosas y musculares, que con frecuencia sufren rápidas entradas de calcio en respuesta a impulsos eléctricos. El sistema de intercambio, inicialmente identificado en músculo cardiaco y nervios de calamar, a diferencia de la ATPasa bombeadora de calcio, no genera de forma independiente la energía necesaria para el transporte de calcio.

Por el contrario, el sistema intercambiador obtiene parte de la energía necesaria para translocar calcio al exterior de la célula de la almacenada en forma de gradiente de concentración transmembranar de sodio. Al igual que el calcio, el sodio abunda más en el medio extracelular que en el intracelular. La concentración de sodio en el fluido extracelular es diez veces superior a la que se da en el interior de la célula, mientras que, para el calcio, la razón de concentración extracelular a concentración intracelular es superior a 1000 por 1. El pequeño gradiente transmembrana de sodio no tiene energía suficiente para expulsar calcio al exterior en contra del empinado gradiente de concentración de este ion.

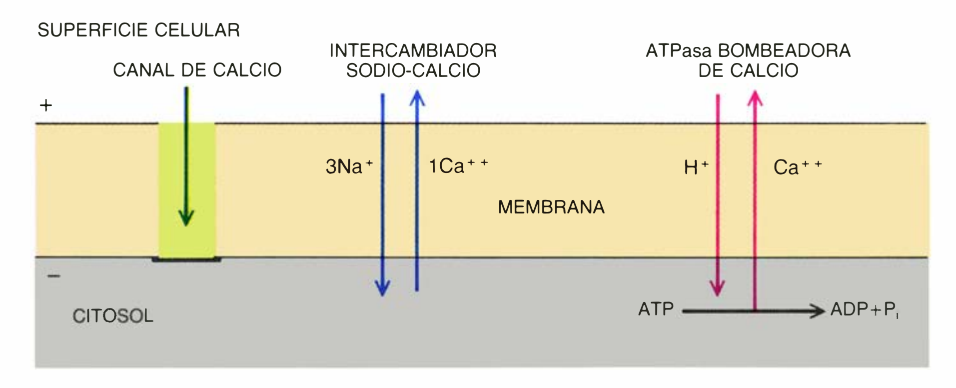
Aun así, se ha demostrado que, por cada ion calcio de doble carga (Ca^{++}) que se expulsa de las células, tres iones sodio (Na^{+}) entran en ellas. Tal desequilibrio de carga permite al sistema de intercambio derivar energía de la almacenada en forma de gradiente de potencial transmembranar: el gradiente de voltaje a través de la membrana de las células excitables. En el estado de reposo, la diferencia de potencial eléctrico transmembranar es de 90 milivolt, siendo el interior negativo respecto del exterior. Dado que el efecto neto de cada uno de los ciclos del sistema de intercambio es el movimiento de una carga positiva al interior de las células, el gradiente eléctrico transmembranar favorece dicho proceso. En consecuencia, ambos gradientes, el químico (gradiente de concentración de sodio) y el eléctrico, suministran energía al sistema de transporte por intercambio sodio-calcio.

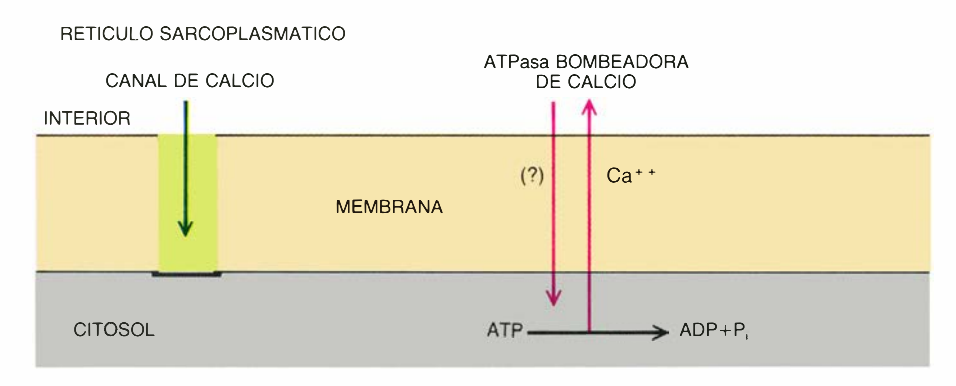
La capacidad del intercambiador es alta: en una célula cardiaca la proteína puede eliminar del citosol hasta 3×10^9 iones calcio por segundo. Sólo lo hace así cuando la concentración intracelular del ion es alta; la velocidad del intercambiador alcanza la mitad de su valor máximo cuando la concentración de calcio es de 1 a 5 micromolar, lo que supone una concentración 10 veces superior a la de reposo. Tales observaciones confirman que la función principal del intercambiador es eliminar grandes cantidades de calcio de células excitables tras haber sido estimuladas.

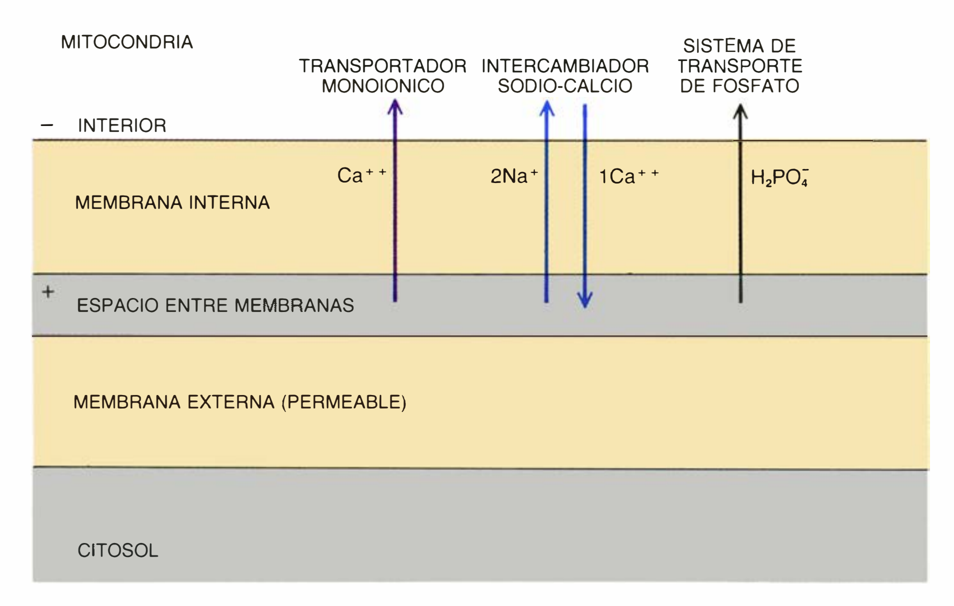
¿Cómo entra el calcio en las células? Tanto en células excitables como en las que no lo son, lo hace a través de canales de calcio: estructuras proteicas que forman poros a través de la membrana y que permiten el movimiento, hacia el interior de la célula, de grandes cantidades de iones a favor de un gradiente de concentración. En las células excitables, los canales de calcio se abren en respuesta a un potencial de acción: una inversión del gradiente de



4. BOMBAS, INTERCAMBIADORES Y CANALES controlan la concentración de calcio en el citosol (fluido citoplasmático) de las células del músculo cardiaco. El calcio entra en las células por canales (*verde*) de la membrana plasmática; abundan esos canales en los túbulos transversos (invaginaciones de la membrana plasmática) y se concentran en las inmediaciones del retículo sarcoplasmático. Cualquier entrada de calcio por los túbulos transversos insta la liberación de ese elemento del retículo sarcoplasmático, probablemente a través de canales de membrana parecidos que se concentran en las terminaciones del retículo. El calcio desencadena la contracción de las miofibrillas musculares. Desde la mitocondria tiene lugar una lenta y continua salida de calcio, mediada por el sistema intercambiador sodio-calcio (*azul*). Expulsan el calcio de la célula una ATPasa bombeadora de calcio (*rojo*) y un intercambiador sodio-calcio. También una ATPasa, localizada en la membrana del retículo, retorna el calcio al interior de ese orgánulo. En la membrana mitocondrial interna, un transportador monoiónico (*púrpura*), cuya translocación del calcio no está compensada por el transporte de otros iones, elimina asimismo calcio del citosol.







5. DETALLES DEL TRANSPORTE DE CALCIO a través de las tres membranas transportadoras de la célula muscular. No se conocen los pormenores de la estructura molecular de las proteínas transportadoras, pero sí se domina su modo de atracción. Los canales de calcio de la membrana plasmática (*arriba*) se abren en respuesta a los cambios del voltaje transmembranar. Permiten estos canales la entrada de calcio en la célula a favor de su gradiente de concentración. El intercambiador sodio-calcio extrae calcio de la célula, dependiendo en parte su ritmo de translocación del gradiente de concentración transmembranar de sodio. Dado que, en cada ciclo, el intercambiador genera un desequilibrio de carga (por cada tres iones de sodio monovalente que entran se expulsa un calcio divalente), el gradiente de voltaje transmembranar aporta energía al sistema de intercambio. La ATPasa que bombea calcio al exterior obtiene energía escindiendo el ATP en difosfato de adenosina (ADP) y fosfato inorgánico (Pi). La membrana del retículo sarcoplasmático (*centro*) parece liberar calcio a través de canales similares a los de la membrana plasmática; liberación que viene inducida por un gradiente de concentración. Una ATPasa bombeadora de calcio obtiene energía del ATP para retornar el calcio al interior del retículo. En la expulsión de calcio a través de la membrana interna de la mitocondria (*abajo*) interviene un intercambiador sodio-calcio, que toma la energía de los gradientes, opuestos, de sodio y calcio. La mitocondria secuestra calcio mediante un sistema de transporte monoiónico, que permite que el ion cargado positivamente se transloque al interior del orgánulo, cargado negativamente. Un tercer transportador de la membrana interna mitocondrial introduce iones fosfato, los cuales reaccionan con el calcio intramitocondrial y forman cristales de hidroxiapatita, la sustancia que mineraliza el hueso.

voltaje transmembranar normal. El potencial de acción se desencadena cuando una molécula mensajera (normalmente un neurotransmisor) se une a su receptor localizado en la superficie de la membrana plasmática de la célula. La unión provoca una despolarización inicial, esto es, un cambio del gradiente de voltaje transmembranar. Seguidamente, los canales de sodio situados en la membrana, que son sensibles al voltaje, se abren, permitiendo la entrada de iones positivos de sodio en el interior de la célula. Con ello, la carga en el lado citoplasmático de la membrana pasa de negativa a positiva.

Los canales de calcio, que normalmente permanecen cerrados, son también sensibles a los cambios de potencial de la membrana. Comienzan a abrirse cuando el potencial alcanza un valor de −30 milivolt; a un potencial de membrana de +30 milivolt, cerca del 70 por ciento de los canales están abiertos. El máximo valor que alcanza el potencial de acción es de unos +40 milivolt. A partir de ese momento, otros procesos iónicos (incluida la salida de iones potasio) retornan el potencial de membrana al valor de reposo, −90 milivolt, y cesa la entrada de iones calcio. Para explicar la sensibilidad de los canales de calcio a los cambios de voltaje transmembranar suele aceptarse que uno de los componentes de la proteína que forma el canal, probablemente constituida por tres, actúa a modo de sensor de voltaje. La proteína presentaría polarización eléctrica, lo que le permitiría reorientarse en la boca del canal atendiendo a las variaciones del campo eléctrico.

En la célula entran unos 3000 iones calcio durante los pocos milisegundos que permanecen abiertos los canales de calcio sensibles al voltaje. Incluso cuando el potencial de acción alcanza su máximo valor, los canales de calcio de la membrana de la célula se abren y cierran continuamente, encontrándose abiertos en cada momento un 70 por ciento de ellos. Puede aumentar significativamente ese porcentaje el AMPc, sintetizado en la membrana plasmática tras la unión de hormonas a sus receptores localizados en la superficie de la membrana, que activa enzimas que fosforilan los canales de calcio. La fosforilación aumenta enormemente la probabilidad de que un canal se abra.

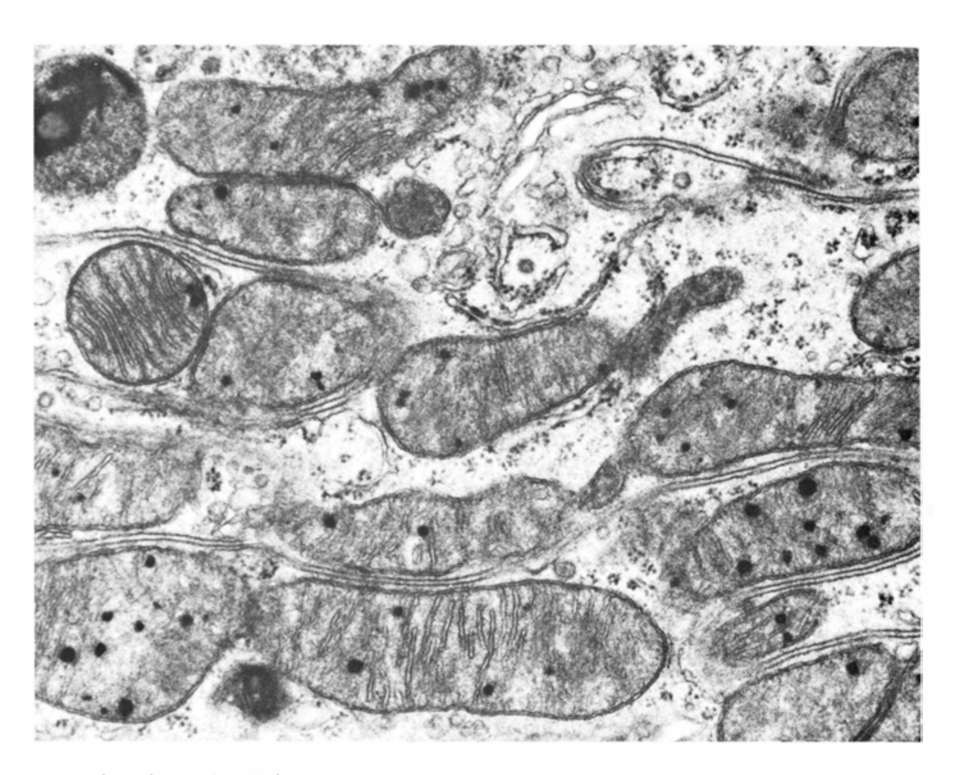
En células no excitables no se producen potenciales de acción, de modo que los canales de calcio de esas células no pueden estar regulados por voltaje.

Quizá se encuentren permanentemente abiertos, o respondan a un control por parte del AMPc, que los abra y cierre gradualmente. En cualquier caso, para evitar grandes entradas de calcio a las células, el número de canales o su conductancia al calcio habría de ser bastante pequeño. Con todo, los canales de calcio de células excitables y no excitables poseen en común diversas propiedades. Los bloqueadores de canales de calcio, drogas ampliamente empleadas en el tratamiento de trastornos cardiacos, inhiben ambos tipos de canales, al igual que lo hace el cobalto que, uniéndose a los receptores de calcio de los canales, los inactiva.

En principio, los mecanismos que translocan calcio a través de la membrana plasmática, en uno u otro sentido, bastarían para controlar la concentración intracelular del ion. En células como las musculares, sin embargo, en las que el calcio desencadena una respuesta rápida y transitoria, la liberación y eliminación del ion debe realizarse con presteza. Un sistema de liberación y recaptación situado en el interior de la célula, en un orgánulo, resultaría más eficaz, por estar más próximo al lugar de acción del calcio. Además, el coste energético del tráfico de calcio en el interior de las células sería inferior al requerido para transportar el calcio entre el citoplasma y el medio extracelular, pues los gradientes de concentración de calcio a través de la membrana de los orgánulos podrían ser menores que el que se da a través de la membrana plasmática.

Se ha medido el área de las membranas transportadoras de calcio, en células de músculo cardiaco, combinándose los resultados con la información disponible acerca de la actividad de las proteínas transportadoras de cada membrana. Los cálculos han revelado que, en músculo cardiaco, el calcio que se introduce o elimina del citosol procede en su mayor parte del trasiego de calcio intracelular, más que del transporte del ion a través de la membrana plasmática.

Dos de los mecanismos implicados en el movimiento intracelular del calcio residen en los retículos endoplasmático y sarcoplasmático. Ambos constan de una malla de túbulos aplanados que delimitan un espacio cerrado en el interior de las células; una de sus funciones es la de secuestrar calcio citosólico. En tejidos no musculares, al retículo endoplasmático le corresponden también otras funciones, como la síntesis proteica y la desintoxicación de productos extraños. En células musculares, su función primordial es la de regular a corto plazo las fluctuaciones de calcio citosólico desencadenantes de la contracción muscular.



6. DEPOSITOS DE CALCIO en forma de hidroxiapatita, recogidos en una micrografía electrónica de mitocondrias de células de riñón de ratón, a un aumento de 30.000 diámetros. Cuando la concentración citosólica de calcio supera notablemente su valor normal, la mitocondria lo absorbe desde el citosol y lo almacena en forma de hidroxiapatita. Aquí, la alta concentración intracelular del ion se consiguió tratando el ratón con hormona paratiroidea, que reduce la excreción renal de calcio. Microfotografía de Feroze N. Ghadially.

Parece probable que los canales de calcio del retículo sarcoplasmático sean análogos a los de la membrana plasmática y que la causa de la apertura de los canales de calcio del retículo sea un primer aumento de la entrada de calcio al citoplasma a través de los canales de la membrana plasmática, abiertos a su vez por un potencial de acción. Los canales de calcio se concentran en unas invaginaciones de la membrana plasmática, denominadas túbulos transversos, que se extienden por el interior de la célula hasta las proximidades de unos ensanchamientos del retículo sarcoplasmático. Por tanto, el flujo inicial de calcio se produce en las inmediaciones del lugar donde ocurrirá su posterior liberación intracelular.

Se desconoce el mecanismo por el cual la entrada de calcio extracelular provoca la liberación del calcio almacenado en el retículo sarcoplasmático de las células musculares. Estudios realizados en otras células han demostrado que un metabolito del fosfatidil-inositol bifosfato (el mismo lípido de membrana que activaba la ATPasa bombeadora de calcio de la membrana plasmática) desencadena la salida de calcio desde el retículo endoplasmático. La unión de hormonas a sus receptores, localizados en la superficie de la membrana, provoca la escisión del lípido. Uno de los fragmentos, el diacil-glicerol, probablemente permanezca en la membrana y, tras combinarse con otro factor, se una al calcio y active ciertas proteína-quinasas con funciones reguladoras. El otro fragmento, el inositol trifosfato, se disuelve en el citoplasma e insta la liberación de calcio intracelular. Se investiga en diversos laboratorios si ese mecanismo opera en el músculo y cuál podría ser su relación con el flujo inicial de calcio, prerrequisito para la contracción muscular.

El calcio que desencadena la contracción muscular lo recapta del retículo sarcoplasmático una ATPasa bombeadora del ion. Al igual que ocurre con la ATPasa de la membrana plasmática, el transportador sarcoplasmático responde rápidamente a pequeños aumentos de la concentración citosólica de calcio, pero difiere de aquélla en muchos aspectos relacionados con su estructura y regulación. La proteína transportadora abunda extraordinariamente en las células musculares. Así, en el músculo esquelético del conejo, por ejemplo, supone hasta un 90 por

ciento del total de proteína de su retículo sarcoplasmático.

Mientras que el retículo sarcoplasmático actúa de controlador rápido y sensible del calcio intracelular, el otro orgánulo que lo regula, la mitocondria, ejerce un control a largo plazo y a gran escala. En la mitocondria se aloja la respiración celular, mediante la cual se obtiene energía a partir de los alimentos. En 1961, Frank D. Vasington y Jerome V. Murphy, de la Universidad Johns Hopkins, descubrieron que las mitocondrias eran también capaces de secuestrar y liberar calcio. Se admite hoy que las mitocondrias no pueden efectuar cambios pequeños (submicromolares) y rápidos de la concentración citosólica de calcio, pero resultan de vital importancia para evitar grandes fluctuaciones del calcio celular, tanto por encima como por debajo de su extensión normal de concentraciones.

Las mitocondrias están delimitadas por una doble membrana. La externa es permeable a los iones, por lo que su papel en el transporte de calcio es pasivo. En contraposición, la entrada de iones de calcio en la mitocondria a través de su membrana interna se vale del gradiente de potencial eléctrico, de unos 180 milivolt, con sentido negativo en el interior de la membrana. El potencial permite que una proteína supuestamente transportadora embebida en la membrana introduzca los iones de calcio hacia la región de carga negativa. El transportador actúa como un portador monoiónico, ya que la translocación de calcio al interior no se compensa por el movimiento de otros iones hacia el citosol.

Mediante el empleo de mitocondrias aisladas y mantenidas bajo condiciones similares a las que se dan en la célula viva, se ha observado que cuando la concentración de calcio en el medio extramitocondrial es del orden de 25 micromolar, la velocidad de captación del ion alcanza la mitad de su valor máximo. Una concentración de calcio de 25 micromolar está muy por encima de su nivel intracelular normal. El transportador monoiónico mitocondrial posee, por tanto, baja afinidad por el calcio. No obstante, su capacidad de transporte es alta: en la mayoría de las células de los organismos superiores, la membrana interna mitocondrial supone el 90 por ciento del total de superficie membranosa con capacidad para transportar calcio.

La considerable capacidad para captar calcio que tiene la mitocondria se compagina con la de almacenamiento. Existen otros transportadores, también localizados en la membrana interna mitocondrial, que translocan fosfato desde el citosol al interior del orgánulo. El fosfato se combina con el calcio formando cristales de hidroxiapatita. De esta manera, la mitocondria puede almacenar grandes cantidades de calcio. Las células normales no suelen almacenar demasiado calcio en la mitocondria, pero una lesión celular a menudo reduce la capacidad de la membrana plasmática para bombear calcio al exterior. Bajo esas condiciones, el citosol se ve invadido por calcio procedente del medio extracelular, y son las mitocondrias las encargadas de secuestrar el exceso del ion.

La capacidad amortiguadora de las mitocondrias no es infinita, y en última instancia puede quedar desbordada su capacidad de almacenamiento si las cantidades de calcio intracelular son excesivas. En esas circunstancias, el calcio, normalmente un elemento vital para el transporte de señales, se convierte en un asesino celular. Las reacciones habitualmente moduladas por calcio discurren de forma ininterrumpida e incontrolada, y el exceso iónico desencadena reacciones que no se dan en las células normales.

En situaciones menos drásticas, durante las cuales las mitocondrias no se ven sobresaturadas, éstas socorren a la célula capturando el exceso del ion. Al cesar la tormenta de calcio, lo liberan en la célula, pero con un ritmo que no trastoca el metabolismo celular. En la liberación interviene un transportador que, al igual que el intercambiador de la membrana plasmática, cataliza un intercambio sodio-calcio. A diferencia del intercambiador de la membrana plasmática, que en cada ciclo creaba un desequilibrio de carga con el intercambio de tres sodios por un calcio, el mitocondrial es eléctricamente neutro: por cada ion calcio que saca del citoplasma entran en el orgánulo dos iones sodio. La neutralidad eléctrica del sistema de intercambio mitocondrial permite que el orgánulo libere el calcio muy lentamente, a pesar del acusado gradiente de voltaje de la membrana mitocondrial, que tendería a acelerar cualquier intercambio eléctricamente desequilibrado. Hay poca o ninguna proteína intercambiadora de sodio-calcio en las mitocondrias de algunas células, por ejemplo, las hepáticas. Se desconoce, en esos casos, la naturaleza del mecanismo que media la liberación del calcio mitocondrial.

En las células de músculo cardiaco el sistema de transporte monoiónico, bombeador del calcio, y el intercambiador de sodio-calcio parecen actuar ininterrumpidamente, si bien a una velocidad muy baja, reciclando el calcio hacia dentro y fuera de la mitocondria de forma continua. Aparentemente, el funcionamiento simultáneo de ambos sistemas de transporte, que se oponen entre sí, constituye un ciclo inútil. El ciclo gasta parte de la energía generada por la respiración, disipando el potencial de membrana mitocondrial. Sin embargo, el funcionamiento continuado de ambos sistemas de transporte asegura que la mitocondria esté, en todo momento, preparada para defender a la célula frente a un aumento súbito y acusado de la concentración citosólica de calcio.

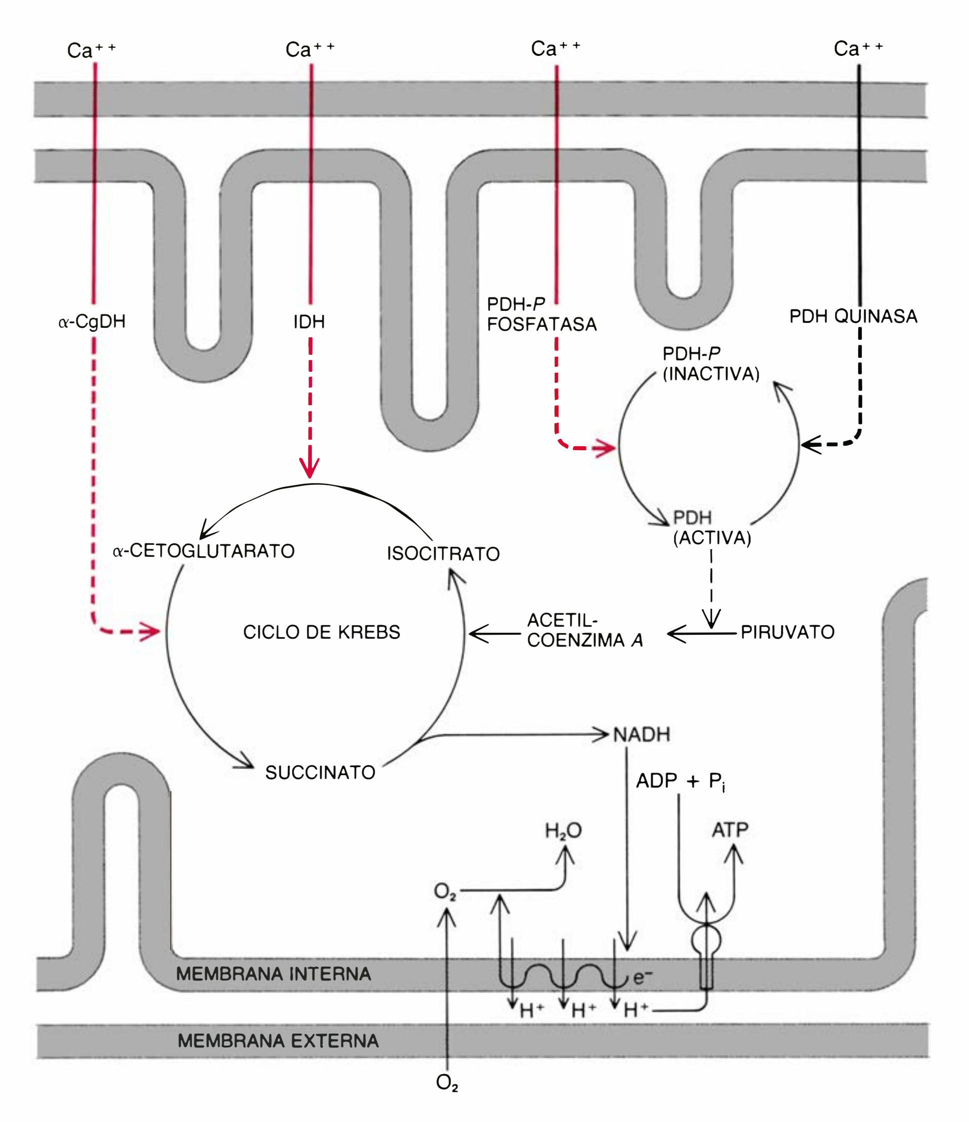
El calcio almacenado en las mitocon-

MEMBRANA (SISTEMAS ELIMINADORES DE CALCIO)	PORCENTAJE DEL TOTAL DE MEMBRANA TRANSPORTADORA DE CALCIO	CONCENTRACION (MICROMOLAR) 0,1	1	10
		PORCENTAJE DEL TOTAL DE CALCIO TRANSPORTADO		
SUPERFICIE CELULAR	0,8			
(ATPasa BOMBEADORA DE CALCIO)		3,6	0,2	0,07
(SISTEMA DE INTERCAMBIO SODIO-CALCIO)		27	3	1,9
RETICULO SARCOPLASMATICO (ATPasa BOMBEADORA DE CALCIO)	12,1	69	90	47
MITOCONDRIA (TRANSPORTADOR MONOIONICO)	87	0	6,4	51

7. COMPARACION DE AREAS Y ACTIVIDADES de las membranas transportadoras de calcio en células de músculo cardiaco. Demuestra que el mecanismo principal por el cual controla la célula la concentración citosólica de calcio es el transporte del ion al interior de los orgánulos, y no la expulsión del calcio a través de la membrana. (En otras células interesa más la expulsión.) Cuando la concentración citosólica es de 0,1 y 1 micromolar (valores típicos de la célula en reposo y de la estimulada, respectivamente), el retículo sarcoplasmático es el principal responsable de la eliminación de calcio. Si la concentración alcanza un valor de 10 micromolar, sólo observado en células dañadas o enfermas, la labor recae sobre las mitocondrias.

drias no es inerte; también le corresponde una función metabólica. Richard M. Denton, de la Universidad de Bristol, ha demostrado recientemente que la actividad de al menos cuatro enzimas mitocondriales se ve afectada por el calcio, cuando la concentración de éste en el orgánulo es del orden de uno micromolar. Catalizan esas enzimas ciertas etapas de la producción de sustancias que transportan a la cadena respiratoria la energía química liberada por la oxidación de nutrientes. La cadena respiratoria, una serie de procesos químicos que se producen a través de la membrana interna mitocondrial, transfiere la energía al portador celular de energía, el ATP. El efecto del calcio es el de acelerar los procesos metabólicos que abastecen la cadena respiratoria y, en consecuencia, aumentar la producción de ATP.

Esta cadena de fenómenos concatenados quizá constituya un mecanismo de retroalimentación que ayude a la célula a defenderse de avalanchas de calcio. Si una célula se ve amenazada por una excesiva concentración citosólica de calcio, debe acudir a todos los mecanismos exportadores del ion. Todos ellos requieren, para su funcionamiento, energía generada por la cadena respiratoria. Las dos ATPasas transportadoras de calcio utilizan el ATP directamente. El intercambiador de sodio y calcio de la membrana plasmática deriva energía de la almacenada en el gradiente electroquímico transmembranar, que a su vez está mantenido por la ATPasa de sodio-potasio. El potencial de membrana mitocondrial, generado por la respiración, aporta energía al transportador monoiónico del orgánulo. A través de sus efectos metabólicos, un aumento inicial de la concentración intramitocondrial del calcio asegura la disponibilidad de la energía requerida para su transporte.



8. EL CALCIO ACELERA LA RESPIRACION, proceso mediante el cual las mitocondrias sintetizan ATP. El calcio estimula (*flechas de color*) una enzima que activa la piruvato deshidrogenasa (PDH); inhibe (*flechas negras gruesas*) otra que inactiva la PDH. El aumento de la concentración de PDH acelera la conversión de piruvato (producto que se obtiene del metabolismo de la glucosa) en acetil-coenzima A, la materia prima que nutre la serie de reacciones conocidas por ciclo de Krebs. Un aumento del contenido de acetil-coenzima A acelera el ciclo de Krebs. Asimismo, el calcio estimula otras dos enzimas que catalizan etapas de dicho ciclo. El resultado neto es un incremento de la cantidad de NADH suministrado a la cadena respiratoria, ubicada en la membrana interna mitocondrial. Es ahí donde la oxidación libera energía para la síntesis de ATP.

Los diferentes mecanismos que regulan el calcio intracelular no actúan por separado. Las redes definen mejor la fisiología celular que las vías aisladas; no constituyen excepción a ello los sistemas intracelulares de señalización. La modulación de los canales de calcio y de la ATPasa bombeadora de calcio por el AMPc constituye un ejemplo de interacción entre dos sistemas de mensajeros intracelulares. La regulación del calcio está también emparejada con la de otros iones: los flujos de sodio y potasio, que inducen el inicio y la finalización del potencial de acción en células excitables, controlan la apertura y cierre de los canales de calcio. A su vez, el aumento de la concentración citosólica de calcio producido tras una serie de potenciales de acción puede abrir los canales de potasio de la membrana plasmática, modificando sus propiedades eléctricas y haciéndola más resistente a subsecuentes despolarizaciones.

Se está desenredando paulatinamente la maraña que constituye la regulación del calcio. Ejemplifica la importancia clínica de esas interacciones una de las drogas de uso clínico más antiguo, la digital. En el siglo XVIII, William Withering, médico inglés, la describió como estimulante de la actividad cardiaca. Hoy se sabe que la digital y otras sustancias afines elevan la concentración intracelular de calcio del músculo cardiaco. No lo hacen interactuando directamente con los transportadores de calcio, sino que inhiben la ATPasa de sodio-potasio de la membrana plasmática, con lo que aumenta la concentración intracelular de sodio. Tal incremento reduce el gradiente de sodio a través de la membrana plasmática, pero lo eleva a través de la mitocondrial. Así, el intercambiador plasmático expulsa poco calcio, mientras que crece el ritmo del intercambiador mitocondrial que libera calcio al citosol. A la postre, se eleva la concentración intracelular de calcio, lo que para muchos pacientes con trastornos cardiacos se traduce en una vida más larga y segura.

Ciencia y sociedad

Nobel de física 1985

Un siglo separa dos experimentos: el llamado efecto Hall del Hall cuántico. En 1879, E. H. Hall, intentando dilucidar si la fuerza que un hilo conductor experimentaba en un campo magnético se ejercía sobre todo el material o sobre lo que ahora llamamos electrones libres, descubrió que cuando una corriente eléctrica (flujo de electrones) pasaba a lo largo de una lámina delgada en una dirección y se aplicaba un campo magnético perpendicularmente a dicha corriente, se establecía un voltaje. Este, llamado en su honor voltaje Hall, producíase en la dirección perpendicular tanto al campo magnético como a la corriente inicial.

En 1980 aparecieron los resultados de un experimento [K. v. Klitzing, G. Dorda and M. Pepper, *Phys. Rev. Letters* **45**, 494 (1980)] donde se mostraba que, en sistemas cuasibidimensionales, la conductividad de Hall (o alternativamente la resistencia de Hall, calculada en el más simple de los circuitos, usando la ley de Ohm, es decir, la razón del voltaje a la corriente) estaba cuantificada. Lo que era lo mismo: sólo adquiría valores discretos en unidades de e^2/h, donde e simbolizaba la carga del electrón y h la constante de Planck. (La unidad de resistencia de Hall sería, pues, h/e^2, unos 25812,80 ohm.)

Esta observación, por la que el alemán K. v. Klitzing ha sido galardonado con el premio Nobel de 1985, exigió un replanteamiento radical de las ideas dominantes sobre transporte bidimensional, especialmente dirigido a desentrañar por qué la presencia de desorden, fuerte desorden incluso, no afectaba al comportamiento "ideal" de dichos sistemas.

Un sistema bidimensional de electrones no existe espontáneamente en la naturaleza, o al menos no lo hemos encontrado, pero puede lograrse en algunos sólidos. Para ello han de darse configuraciones especiales. La habitual es la conocida por "efecto de campo metal-óxido-semiconductor" (MOSFET, por sus siglas en inglés). Consiste en un "sandwich" formado por una lámina aislante (óxido metálico) con un metal y un semiconductor en cada lado. Bajo las condiciones apropiadas, la aplicación de un campo eléctrico curva la banda de conducción del semiconductor y atrae los portadores de carga hacia la "interfaz" del semiconductor-aislante. Cuando la región resultante de esa curvatura de las bandas es muy estrecha, de unos 30 angstrom, los niveles de energía en la dirección perpendicular a la superficie están cuantificados y separados por valores del orden de 30 meV. Por haberlo sugerido, Schrieffer compartió con Cooper y Bardeen el premio Nobel de física de 1972.

Si esta diferencia de energía es mayor que las relevantes del problema, el movimiento en la dirección perpendicular a la superficie está congelado y el sistema se comporta como un gas de electrones bidimensional. Peculiaridad interesante de dichos sistemas es que la densidad de portadores puede controlarse de forma continua mediante el potencial aplicado. El gas de electrones libre bidimensional presenta unos rasgos propios, fácilmente deducibles con un conocimiento elemental de las reglas de la mecánica cuántica.

En dos dimensiones, el número de estados permitidos por unidad de área, $d_0=(m/2\ h^2)$, a una determinada energía E, no depende de la energía. (En claro contraste con la densidad de estados en tres dimensiones donde el número de estados por unidad de volumen aumenta con la energía.) En dos dimensiones, la resistencia coincide con la resistividad y es independiente del tamaño de la muestra; sólo depende de su forma. Al aplicar un campo magnético a un gas de electrones libres, el cuasicontinuo de niveles de energía existente en ausencia de dicho campo se condensa en un conjunto de niveles discretos, llamados niveles de Landau, separados por la energía asociada a la frecuencia de la órbita clásica de un electrón en un campo magnético B. (Expresado en forma técnica: $hw_c = eH/mc$; w_c se conoce con el nombre de frecuencia de ciclotrón.) El número de electrones existente en un nivel de Landau depende del campo magnético y viene dado por la expresión $d_0hw_c=eH/hc$; la degeneración (número de electrones de cada nivel) aumenta linealmente con el campo.

En un sistema bidimensional ideal, la explicación del efecto Hall cuántico no reviste mayor dificultad. La conductividad de Hall se define como nec/H, donde n representa el número de portadores de carga. Si el número de niveles de Landau por debajo de la energía de Fermi es N, el número de portadores viene dado por NeH/hc, por lo que la conductividad de Hall será Ne^2/h. Sólo dependerá de constantes de la naturaleza.

Esta simple explicación, que se infiere del comportamiento ideal de un gas de electrones libres, no puede extenderse a un sistema real donde los electrones "chocan" con otros electrones y "sienten" las desviaciones de la perfecta periodicidad; de hecho, el desorden es inherente al efecto Hall. Si no hubiese desorden, no habría estados en el intervalo ("gap") entre los niveles de Landau, ni el nivel de Fermi podría estar en el intervalo para un rango finito del campo magnético o del voltaje aplicado.

La paradoja estriba en que el efecto Hall cuántico no puede observarse sin desorden y, sin embargo, no hay ninguna razón obvia por la que en vista del sencillo argumento presentado exista con desorden. Efectivamente, las colisiones asociadas al desorden hacen que los niveles de Landau se presenten ensanchados en energía. Pero ahora los niveles, en vez de presentarse de forma discreta sin ningún estado posible entre ellos, se ensanchan y hay estados permitidos a energías situadas entre los niveles originales. Por ello, y de acuerdo con el argumento expuesto antes, la variación de la conductividad no sería escalonada sino que debería presentar cierta continuidad. Lo sorprendente del resultado experimental es que las mesetas ("plateau") aparecen perfectamente planas, según cabría esperar de un gas ideal, a pesar del desorden presente. La disolución de esta paradoja arranca del fenómeno de localización electrónica debida al desorden, o localización de Anderson.

En un sólido ideal, de periodicidad perfecta, las colisiones elásticas de los electrones por los potenciales iónicos no se traduce en ninguna resistencia al movimiento de los electrones, y éstos se propagan con libertad por el sólido. Se trata del denominado teorema de Bloch: los estados electrónicos se extienden por todo el material. Phil Anderson demostró en 1958 que si el desorden era suficientemente fuerte, no podían darse estados electrónicos extendidos y los electrones se hallarían en estados localizados; vale decir: sus funciones de onda decrecen exponencialmente con la distancia, en vez de extenderse por todo el espacio.

La idea de Anderson fue desarrollada posteriormente por Mott [ver *Investigación y Ciencia*, enero de 1978],

quien demostró que, aun cuando el desorden no era suficientemente grande para localizar todos los estados, había una energía crítica (llamada "mobility edge"), que separaba estados extendidos y localizados.

Cabría esperar, a primera vista, que el número de portadores de carga que contribuyeran al efecto Hall fuese menor que en el caso ideal, invalidando la cuantificación. Sin embargo, lo que experimentalmente se detecta es una perfecta cuantificación debida, se supone, a que los estados no localizados intervienen de una manera más destacada en la conductividad, cancelando exactamente el defecto creado por los que se localizan. Una explicación de este fenómeno, desde una óptica muy original, ha sido sugerida por Laughlin al proponer que la cuantificación es exacta merced, sobre todo, a la invarianza de aforo o gauge.

Tsui, Stormer y Gossard acaban de mostrar la existencia de cuantificación fraccional: N=1/3 ; 2/3. Dichos estados ocurren cuando el estado de mínina energía está fraccionalmente ocupado, debido, se cree, a la existencia de un estado fundamental correlacionado e inducido por la interacción de Coulomb.

El efecto Hall cuántico prueba que la física, por encima de todo, es una ciencia experimental y que los argumentos basados en belleza, coherencia y elegancia, sin negar su incidencia, son al mismo tiempo subjetivos y, en cualquier caso, han de someterse al test del experimento. El descubrimiento experimental del efecto Hall cuántico ha obligado a un serio replanteamiento del problema del transporte en sistemas bidimensionales.

Añádase, por último, que el efecto Hall cuántico puede resultar de gran utilidad para determinar con mayor exactitud la constante de estructura fina, así como para definir una resistencia estándar. La constante de estructura fina es adimensional y mide el acoplamiento entre la materia y la radiación, el campo electromagnético. Desempeña un papel central en la teoría de la electrodinámica cuántica (QED). Dicha constante se relaciona con la resistencia Hall a través de un factor de proporcionalidad en el que interviene la permeabilidad del vacío, que se define de forma precisa, y la velocidad de la luz en el vacío, que se conoce con gran exactitud, razón por la cual se usa el efecto Hall cuántico para determinar la constante de estructura fina. Una mayor precisión de nuestro conocimiento de la constante de estructura fina serviría para corroborar la coherencia interna de la electrodinámica cuántica.

Como hemos visto, la cantidad h/e^2 tiene unidades de resistencia eléctrica, con un valor que se encuentra en un rango adecuado para medidas de 25812,80 ohm. Si la resistencia Hall se revela totalmente independiente de la muestra empleada en su medida puede usarse como estándar, en la que la unidad de resistencia aparece determinada en función de constantes fundamentales. Una gran ventaja sería la de facilitar la comparación internacional de estándares de resistencia. (Pedro Miguel Echenique.)

Marcadores y tumores

Los tumores germinales son neoformaciones malignas que se originan a partir de las células germinales de las gónadas (testes y ovario). A pesar de su baja incidencia, representan el cáncer más frecuente del varón en el período crítico comprendido entre los 20 y 34 años. Dos atributos que los caracteriza es el crecimiento rápido y su pronta diseminación fuera de los límites de la gónada, que contribuyen al fallo de terapéuticas de índole local.

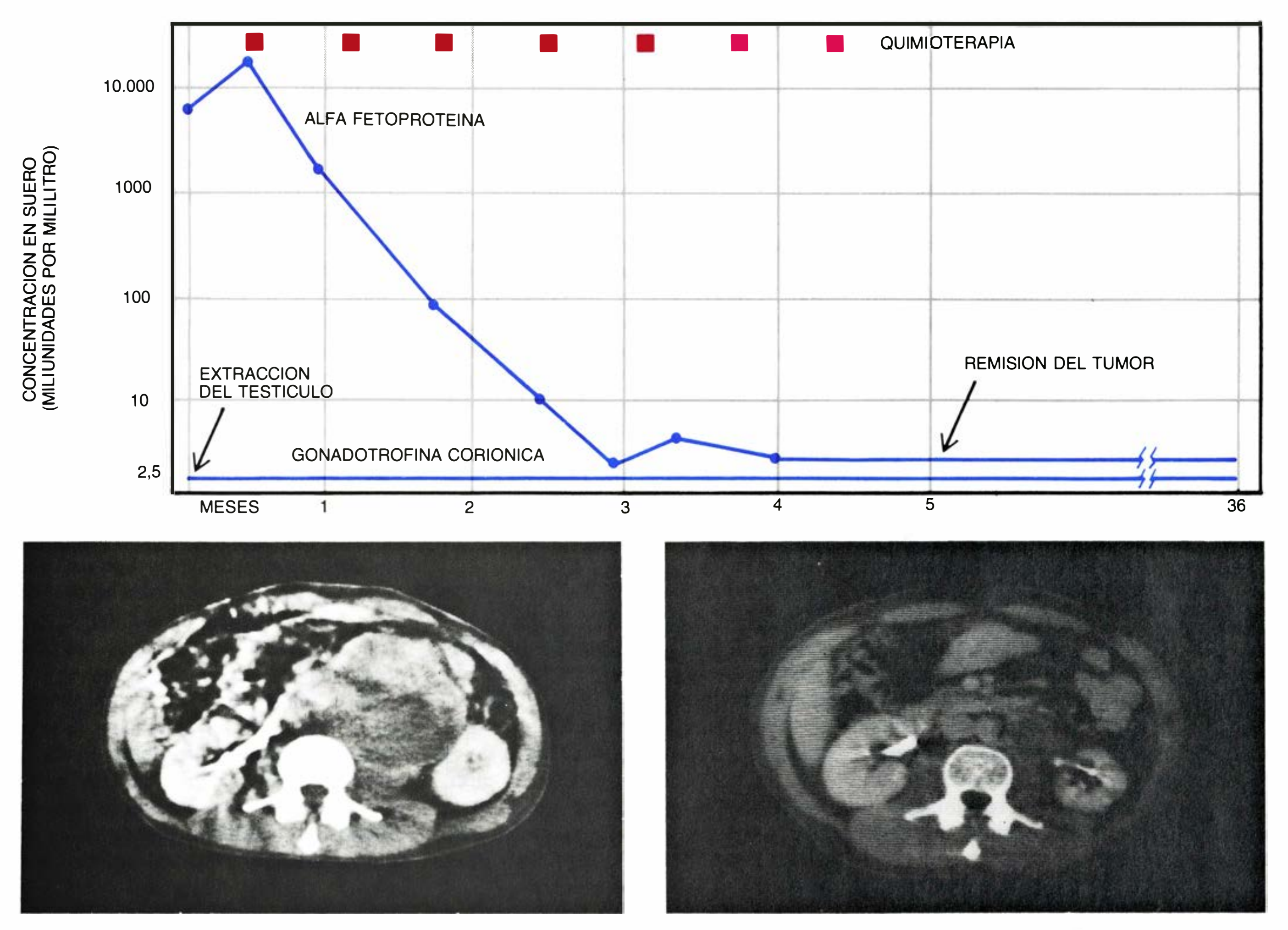
Tales circunstancias explicaban la pérdida inexorable de los pacientes poseedores de tumores con metástasis (fracciones de tumor que anidan fuera de la contigüidad con el mismo) hasta que, en 1977, Larry Einhorn y John Donohue, de la Universidad de Indiana, obtenían por primera vez el 100 por ciento de respuestas en enfermos con cáncer de testículo diseminado, con una combinación de tres fármacos anticancerosos, en la que incluían un derivado del platino (cis-dicloro-diamino-platino) de potente acción frente a este tipo de carcinomas. Estudios subsiguientes mostraron que, de cada tres, dos pacientes que conseguían una respuesta clínica completa podían considerarse definitivamente curados.

Si bien hay que imputar al empleo de las nuevas asociaciones de drogas el papel más relevante en la consecución de estas curaciones, otro logro brilla con luz propia dentro de la obtención de tal avance: el uso racional y sistemático de los marcadores tumorales alfa-1-fetoproteína y gonadotrofina coriónica. Un marcador tumoral ideal sería la sustancia que, detectada en cualquiera de los líquidos orgánicos de un individuo, evidenciara la existencia de una neoformación maligna. Su concentración debería estar en íntima correlación con el volumen total de tumor y su propia naturaleza debería señalar su lugar de origen dentro del cuerpo. Aunque se hable de un centenar largo de sustancias como marcadores tumorales, sólo media docena ha demostrado su utilidad en el empleo clínico diario. Entre ellas, alfa-1-fetoproteína y gonadotrofina coriónica.

La primera de estas sustancias glicoproteicas se sintetiza en las células del saco vitelino e hígado del embrión, desapareciendo al mes del nacimiento, mientras que la segunda se origina en las células del sincitiotrofoblasto de la placenta con la función de preservar la secreción de la progesterona responsable de mantener el embarazo en sus fases iniciales críticas. La producción de estos dos marcadores tumorales por parte de la mayoría de tumores germinales los ha convertido en un modelo ideal para investigar la utilidad clínica de la determinación de estas sustancias en el cáncer.

En el diagnóstico relativamente tardío ha encontrado la medicina su peor obstáculo. Como ha demostrado el grupo de Bosl, de la Universidad de Minnesota, escasas semanas de retardo diagnóstico significan la extensión del tumor germinal fuera del ámbito local con el empeoramiento del pronóstico consiguiente. A resolver esa dificultad se han encaminado los trabajos del grupo de T. A. Waldman, del Instituto Nacional de Cáncer (NCI) norteamericano en Bethesda, y el del propio autor en el Hospital de la Santa Cruz y San Pablo (Barcelona). Se ha demostrado ya que más del 80 por ciento de las masas testiculares cuya naturaleza es un tumor germinal no-seminoma (variante histológica más agresiva) presentaban elevaciones de los marcadores en plasma antes de la orquiectomía, mientras que los niveles permanecían indetectables en las masas benignas. Ello supondría la detección y tratamiento precoz de los pacientes con peor pronóstico.

No todos los tumores germinales son capaces de producir marcadores. Robert J. Kurman, del Instituto de Patología de las Fuerzas Armadas en Washington, en colaboración con el grupo del NCI antes citado, localizaron, mediante una técnica de tinción indirecta con inmunoperoxidasas, las células responsables de la producción de marcadores, atribuyendo la secreción de alfa-1-fetoproteína a células del seno endodérmico y de gonadotrofina coriónica a células gigantes del sincitotrofoblasto, lo que explicaba que tumores carentes de estos tipos celulares no se-



*1. Estrecha correlación entre el descenso y ulterior negativización de lo marcadores tumorales (*gráfica*) y la reducción de una metástasis retroperitoneal (*T*) de gran tamaño (*abajo, izquierda y derecha*). Tras la extirpación del testículo se sometió el paciente a un tratamiento quimioterapéutico (*cuadrados de color*). A intervalos de tres semanas recibió, en cinco ocasiones, cis-platino, vimblastina y bleomicina (*color intenso*) y, en dos ocasiones más, cis-platino, adriamicina y etopósido (*color claro*). El análisis posterior a la extirpación de la masa tumoral residual mostró que ésta no constaba más que de tejido tumoral destruido*

cretaran marcador. Recientemente, nuestro grupo ha demostrado que, con la nueva determinación en marcadores en plasma, es posible detectar un 25 por ciento de pacientes con dichas variedades histológicas que han pasado desapercibidas en el dictamen histopatológico rutinario y cuya existencia es decisiva para elaborar un pronóstico.

Es en el extenso apartado del seguimiento de los pacientes tras la extirpación inicial de la gónada enferma donde los marcadores han mostrado su utilidad máxima. Eficacia que radica, fundamentalmente, en la capacidad de los marcadores de detectar volúmenes tumorales mínimos, no registrados por ninguna otra técnica radiológica ni de laboratorio. Ello permite iniciar el tratamiento con quimioterapia anticancerosa en estadios más tempranos de la historia natural del tumor, cuando éste ha tenido menores probabilidades de metastatizar o de sufrir mutaciones espontáneas que condicionen clones celulares resistentes a dichos fármacos, tal como han señalado en sus recientes trabajos J. H. Goldie y A. J. Coldman, de la Universidad de Columbia Británica de Canadá. Esta presunción queda respaldada por las tasas de curaciones absolutas conseguidas en pacientes tratados con quimioterapia, cuando niveles altos de marcador son el único reflejo clínico de la existencia de cáncer en actividad.

En el tratamiento de un paciente neoplásico importa conocer cuanto antes la eficacia de la terapéutica administrada. La evaluación de la respuesta a un tratamiento a través de la exploración física y técnicas radiológicas complementarias se resiente del largo intervalo de tiempo que debe transcurrir para identificar clínicamente dicha respuesta. Los trabajos pioneros de K. D. Bagshawe, del Hospital Charing Cross de Londres, y N. Javadpour, del NCI, han demostrado que la determinación seriada de los marcadores constituye el método más rápido y fidedigno de seguir el desarrollo de la respuesta a un determinado tratamiento.

Además de haberse convertido en imprescindible para elaborar la estrategia terapéutica a realizar en cada instante, la determinación de marcadores permite precisar en qué momento debe detenerse el tratamiento quimioterápico por haberse destruido, en teoría, el mal. Frente a opiniones extendidas de que el número de ciclos de fármacos a administrar debe ser invariable para la mayoría de enfermos, nuestro grupo del Hospital de la Santa Cruz y San Pablo y el de K. D. Bagshawe y R. Begent, en el Hospital Charing Cross de Londres, basamos la duración de la terapéutica en la sensibilidad determinada de cada tumor, que queda reflejada por la velocidad de descenso de los marcadores tras cada ciclo de qui-

mioterapia. Así, pacientes con descensos drásticos de sus marcadores recibirían menos cantidad de quimioterapia que aquellos otros con descensos más lentos. Dicha estrategia nos ha permitido obtener cifras de un 80 por ciento de curaciones acortando considerablemente la duración del tratamiento.

La tipificación precoz de aquellos enfermos que no conseguirán la remisión completa, a pesar de recibir la terapéutica tan eficaz ya mencionada, ha sido la principal preocupación del autor de este escrito. Primero, en colaboración con R. Begent y K. D. Bagshawe, en el Hospital Charing Cross de Londres, y posteriormente con mi actual grupo de trabajo hemos contribuido a demostrar que niveles muy altos de alfa-1-fetoproteína y gonadotrofina coriónica son el factor pronóstico más ominoso en pacientes con tumores germinales gonadales, incluso por encima del mero volumen tumoral. Ello significa que la dosificación del marcador antes de iniciar el tratamiento permite determinar el grupo de pacientes en quienes deben ensayarse estrategias terapéuticas diferentes de las habitualmente empleadas. R. F. Ozols y colaboradores, del NCI, D. Vugrin, del Hospital Memorial de Nueva York, y M. Peckham están ofreciendo en la actualidad resultados esperanzadores con nuevas terapéuticas más agresivas en estos pacientes de mal pronóstico. Así las cosas, se puede afirmar que el binomio tumor germinal-marcador tumoral es una de la simbiosis más perfectas del mundo oncológico, y que ha contribuido de forma determinante en la consecución del mayor avance registrado en la curación del cáncer de los últimos años. (José Ramón Germà Lluch.)

El reparto del pastel

Por fin, el mes de junio del año pasado se hizo público el informe elaborado en Gran Bretaña por la Comisión Kendrew, creada en marzo de 1984 a instancias de los órganos responsables de la financiación de la ciencia y la tecnología para analizar la participación británica en la física de altas energías y en particular en el CERN (Centro Europeo de Investigación Nuclear). Tras más de un año de trabajo, la Comisión, presidida por el biólogo molecular Sir John Kendrew, con participación de otros expertos, aunque ninguno del campo que se examinaba, presentó un extenso análisis.

El informe, desde su publicación, ha recibido críticas de muchos sectores, desde la prensa diaria hasta las revistas científicas *Nature, New Scientist* y otras. Se recrimina la incoherencia entre el cuerpo del informe y sus conclusiones. En efecto, tras dedicar unos elogios difícilmente superables al nivel de la física de altas energías alcanzado por los investigadores británicos y a los beneficios de su participación en el CERN, y después de poner como modelo de cooperación científica europea dicho centro, el informe acaba proponiendo que, en el momento en que se concluya el nuevo acelerador LEP ("Large Electron Positron") en 1989, se deberá reducir la contribución británica en el presupuesto del CERN, de manera que en 1991 tal asignación quede recortada en un 25 por ciento. Aparte la incoherencia que supone economizar en algo ventajoso, tal recorte resultará fatal para la física británica, pues indica que, a pesar del esfuerzo de construir el LEP, no se participará en los experimentos que en dicho colisionador se realicen; esa cuarta parte de menos incidirá principalmente en la realización de los experimentos.

Es probable que la creación de la Comisión haya surgido de la propia comunidad científica que, en una atmósfera de recorte de los fondos de investigación, pensaba que la reducción de la cuota del CERN redundaría en beneficio de la propia investigación. Pero ahora es la propia comunidad científica la que está alarmada y se da cuenta de que este tipo de rencillas entre sectores no beneficia a nadie. Por descontado que los físicos de partículas elementales se han puesto en pie de guerra, y si bien están de acuerdo con los controles que garanticen el nivel y el rendimiento de la investigación, se preguntan, como hizo el profesor Abdus Salam en una reciente reunión en el Imperial College, si no "debería haber un informe Kendrew sobre la investigación en Defensa, sobre su calidad y rendimiento", ya que en este caso no se trata de un pequeño porcentaje del PIB, sino de una fracción sustancialmente mayor.

El caso británico puede también tener su moraleja en otros países, España incluida. Más de una vez se producen roces entre distintos sectores de nuestra débil comunidad científica porque parece que unos reciben más que otros. Hay que tener en cuenta que la única manera de que la ciencia progrese es mediante un esfuerzo común hacia adelante de todos los sectores. La obtención de una mejor financiación de la investigación no vendrá de distintas maneras de repartir el pastel, sino que hay que lograr que éste sea mayor. (R. Pascual.)

Retazos litosféricos

Son bloques de corteza limitados por fallas y yuxtapuestos a los antiguos núcleos de los continentes. Su acreción determina el aumento de extensión de los continentes y su remodelamiento en lo que vienen a ser mosaicos geológicos

David G. Howell

Desde hace más de un siglo los geólogos vienen tratando de descifrar el motor que produce los grandes accidentes geológicos de la Tierra. Los primeros investigadores concibieron ciclos geosinclinales: grandes combamientos corticales que se rellenan de sedimentos, seguidos de levantamientos que alzan cordilleras jóvenes. Esa interpretación dio paso a la teoría de la tectónica de placas, planteada en la década de 1960, según la cual las direcciones de movimiento son predominantemente horizontales: la capa externa, frágil, de la Tierra consta de grandes placas corticales que se mueven sin cesar. Donde las placas se alejan una de otra se abren valles de fractura, o rifts, y se forman nuevas cuencas oceánicas; donde las placas chocan se levantan cadenas de volcanes según líneas paralelas a la zona de colisión; donde resbalan una frente a otra, siguiendo fallas como la californiana de San Andrés, suelen registrarse terremotos de gran intensidad.

Con todo, parece hoy necesaria una nueva revisión de esa teoría. Los modelos corticales engendrados por actividad de tectónica de placas son efímeros: fuerzas tectónicas reciclan los patrones originarios, troceando fragmentos de corteza, dispersándolos y remodelándolos en agrupamientos de bloques corticales dispares. Simultáneamente, de los procesos volcánicos surgen bloques corticales nuevos, que se incorporan al reciclaje. Así, las placas corticales vienen a ser una extraña mezcla de retazos de fragmentos corticales, mosaicos geológicos montados a partir de retazos litosféricos, denominados también litosferoclastos.

El concepto de retazo litosférico, o litosferoclasto, surgió en la década de 1970 cuando, a raíz de ciertos conflictos sobre derechos de uso del territorio en Alaska, el Servicio Geológico de los Estados Unidos envió equipos de geólogos a explorar los recursos minerales. Sus hallazgos resultaron asombrosos. La elucidación de un patrón geológico en una porción del estado inducía a predecir el patrón previsible pocas decenas de kilómetros más allá. Y, sin embargo, el patrón real resultó marcadamente diferente: la roca tenía otra edad y otra composición. Resumiendo, la aplicación directa de la teoría de la tectónica de placas no explicaba la geología de Alaska; según se comprobó, el estado venía a ser una aglomeración de fragmentos corticales. Alaska es los restos corticales del antiguo océano desaparecido que precedió al Pacífico. Es un mosaico de litosferoclastos desmembrados y reubicados durante los últimos 160 millones de años a raíz de las migraciones y choques de placas corticales. Todavía siguen llegando pedazos desde el sur.

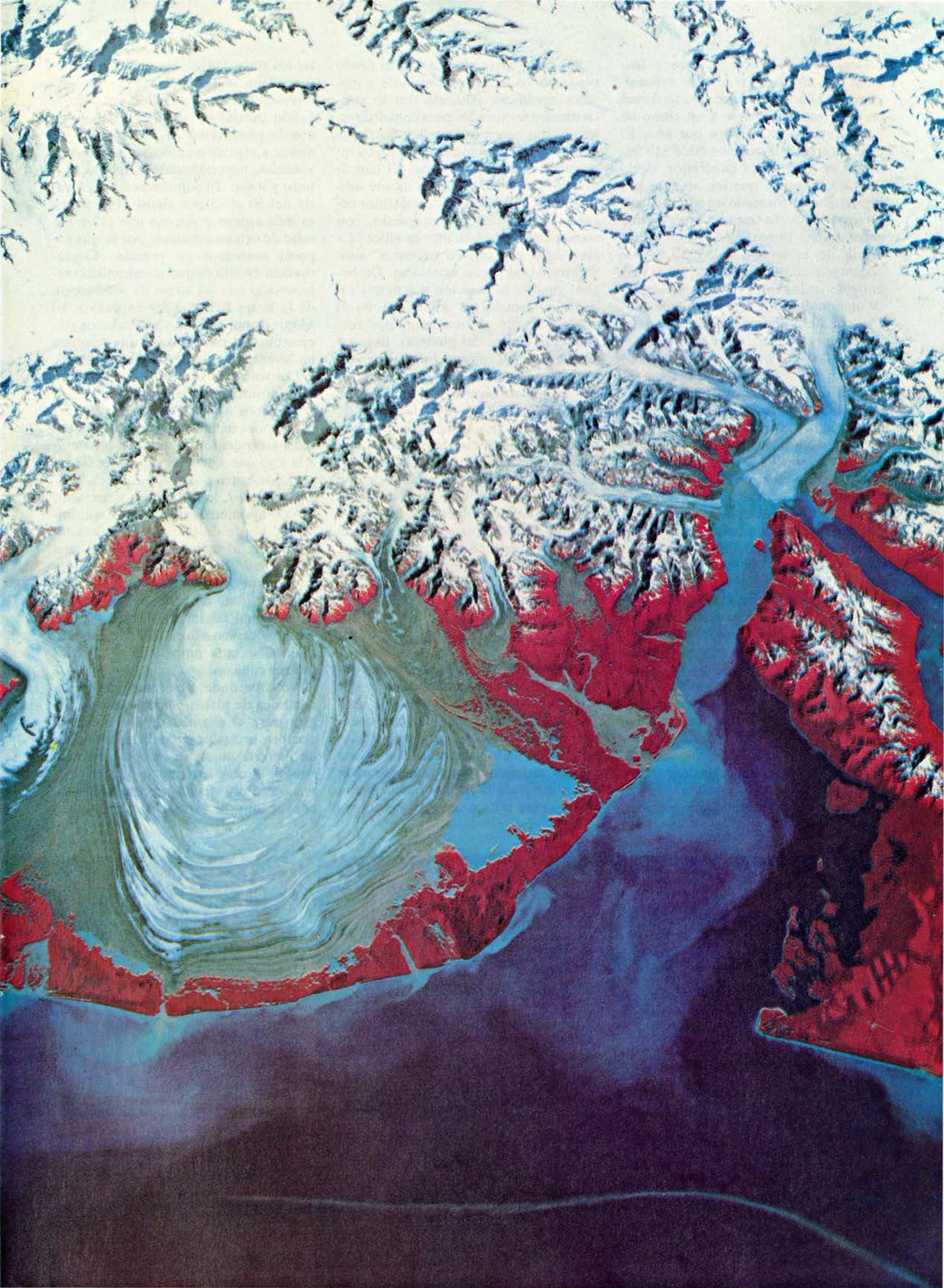
Formulación de la tectónica de placas

Un breve repaso a la teoría de la tectónica de placas ayuda a situar en el marco adecuado los retazos litosféricos. En última instancia, no representan más que el aspecto más moderno del esfuerzo central de la geología, un esfuerzo por intuir la vastedad del tiempo y comprender los efectos acumulativos de los movimientos lentos en la tierra. (La mayoría de los procesos tectónicos se desarrollan a velocidades iguales o inferiores a la del crecimiento de las uñas.) Fundamentalmente, en la corteza terrestre se distinguen dos dominios: la corteza oceánica, densa y homogénea, y los continentes, más livianos y mineralógicamente heterogéneos. Los litosferoclastos son los fragmentos corticales incorporados a los antiguos núcleos de los continentes. Describiendo su historia se enfrenta uno a todas las consecuencias de la teoría de la tectónica de placas, según la cual la corteza oceánica transporta los continentes cual si fuera una enorme y lenta cinta transportadora.

Los océanos constituyen ciertamente una pieza central de la tectónica de placas. En particular, las cuencas oceánicas se ensanchan coincidiendo con la divergencia de las placas resultante de la ascensión y solidificación de magma, roca fundida, procesos éstos que crean corteza oceánica a lo largo de las prominencias submarinas, o dorsales, denominadas centros de expansión oceánica. En el transcurso del tiempo geológico puede que esos procesos se frenen, reduciendo su extensión la malla de centros expansivos. La longitud actual de esa malla se cifra en unos 56.000 kilómetros, y durante los dos últimos miles de millones de años las velocidades medias de expansión probablemente hayan rondado los cinco centímetros anuales. El Atlántico de hoy se expande a menos de tres centímetros por año y las partes más activas del Pacífico Oriental lo hacen a unos 16 centímetros anuales. (En estos ritmos se combinan las velocidades de las dos placas que se alejan en sentido opuesto desde un centro de expansión oceánica.)

La multiplicación de ambas cifras

1. SEIS LITOSFEROCLASTOS ocupan la parte costera de Alaska sudoriental en esta imagen Landsat. La bahía Yakutat indenta la costa; a su oeste, el glaciar Malaspina desciende desde un grupo de picos entre los que se cuentan el monte San Elías y el monte Augusta. La roca subyacente al hielo y la nieve corresponde a partes de islas volcánicas, partes de un margen continental desplazado o los productos metamorfoseados y refundidos de sedimentos en una matriz sedimentaria. En cada caso, la roca acreció hasta el antiguo núcleo de Norteamérica durante los últimos 100 millones de años: los litosferoclastos son fragmentos dispares de corteza, barridos juntos por los movimientos de grandes placas litosféricas de la tierra. En Alaska meridional, los litosferoclastos son cuerpos alargados resultantes de la rebanadura de la corteza por fallas meridionales. Incluso hoy, el fallamiento sigue esparciendo litosferoclastos hacia el norte. Atrapados entre la placa norteamericana y la placa pacífica, los litosferoclastos de Alaska meridional han girado en sentido horario.



(velocidad media de expansión y longitud del sistema de centros expansivos) permite estimar que hoy se forma corteza oceánica nueva a un ritmo de 2,8 kilómetros cuadrados por año. El área cubierta por océanos mide 310 millones de kilómetros cuadrados, de lo que se desprende que los océanos podrían haberse formado en sólo 110 millones de años. Se trata sin duda de una constatación importante. Antes de la teoría de la tectónica de placas los océanos se consideraban las partes más antiguas de la tierra, puesto que son las de menor altitud. El razonamiento subyacente a esa interpretación era que la roca vieja está más fría y, por tanto, es más densa que la roca joven, situándose por debajo de ésta. La notoria juventud de los océanos la han confirmado plenamente las muestras de corteza oceánica recogidas por el Proyecto de Perforación del Océano Profundo (*Deep Sea Drilling Project*). En la tierra actual, la edad de la corteza oceánica va desde cero años, a lo largo de las crestas de las dorsales submarinas que señalan los centros de expansión, hasta no más de 180 millones de años, en el Pacífico Oriental, la parte de fondo oceánico que más distante se halla de una dorsal. En los últimos 2000 millones de años pueden haberse creado y destruido hasta 20 océanos.

La corteza oceánica asciende en una dorsal, se desplaza a todo lo ancho de un océano y se hunde en una fosa, que señala lo que se ha dado en llamar zona de subducción. No por ello la superficie de la corteza es plana. A grandes rasgos, la roca oceánica se enfría y se comprime a medida que se aleja de la cresta de la dorsal donde se emplazó en la corteza, hundiéndose paulatinamente.

Por otra parte, abundan en el fondo oceánico los montes submarinos y mesetas oceánicas. Muchos son lo suficientemente elevados para constituir islas. Como veremos, pueden barrerlos los procesos de tectónica de placas e integrarse en litosferoclastos. El constituyente fundamental de un monte submarino es el basalto, roca volcánica oscura, rica en hierro y magnesio, con menos del 50 por ciento de sílice. La roca sube en "puntos calientes" subyacentes a las placas oceánicas. De hecho, cuando la posición del punto caliente es estable (es decir, cuando el punto caliente se mantiene fijo respecto del núcleo del planeta), llegan a formarse largas cadenas lineales de volcanes a medida que la placa avanza sobre el chorro ascendente de magma. Las islas Hawai forman parte de una de esas cadenas. La velocidad de crecimiento de la cadena es bastante alta comparada con lo que es habitual en los procesos geológicos, a pesar de que el propio punto caliente raramente mide más de un kilómetro de diámetro. La contribución planetaria anual del amontonamiento de basalto oceánico al aumento de la corteza continental se estima en sólo 0,2 kilómetros cúbicos.

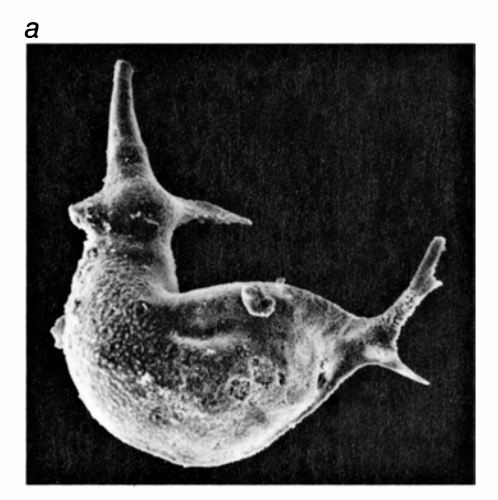
Vulcanismo de zona de subducción

Mayor contribución a la corteza continental se produce en el choque de dos placas. A lo largo de la zona de choque, la placa más densa desciende, encontrándose rodeada de temperaturas progresivamente mayores. La placa acarrea sedimento y agua aprisionada en los poros de la roca cortical. A una profundidad crítica, generalmente entre los 100 y 150 kilómetros, el agua desencadena una secuencia de acontecimientos físicos y químicos, incluida la fusión parcial de roca, que culminan con la producción de un magma que tiende a ser rico en elementos químicos volátiles, especialmente aluminio, potasio y sodio. El contenido en sílice varía del 50 al 75 por ciento. El magma es más espeso y viscoso que el de basalto de origen oceánico, por lo que soporta aumentos de presión. Consecuencia de ello es que el vulcanismo relacionado con las zonas de subducción de la tierra tiende a ser explosivo. El Monte Santa Helena y el Krakatoa son ejemplos de ese proceso, que también ha levantado muchos arcos insulares.

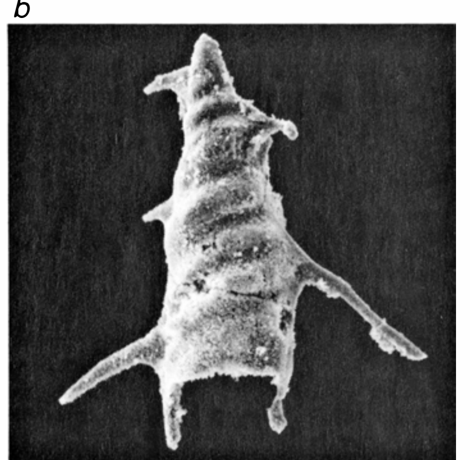
Los volcanes de borde de placa difieren considerablemente de los montes submarinos e islas del interior de una placa, pues se hallan sobre grandes cortinas ascendentes de magma, paralelas a la fosa que señala la superficie de separación entre dos placas que chocan. En conjunto, la tierra hoy tiene unos 37.000 kilómetros de cadenas volcánicas de borde de placa; por cada kilómetro de tal actividad se calcula una erupción de 20 a 40 kilómetros cúbicos de material silíceo nuevo en un millón de años. Así, en todo el planeta, el ritmo anual de incorporación de material volcánico silíceo a la corteza continental se cifra entre 0,75 y 1,5 kilómetros cúbicos.

No cabe duda, por tanto, de que la tectónica de placas brinda una base lógica para el examen de los litosferoclastos y la clasificación de las adiciones de material nuevo a la corteza continental. Probablemente la propia corteza oceánica apenas aporte material, por cuanto la formación de corteza

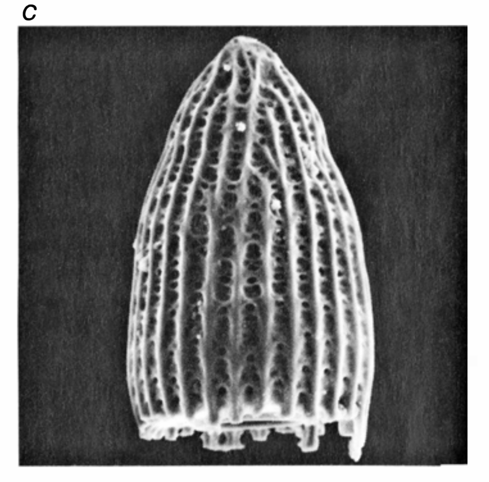
a



b



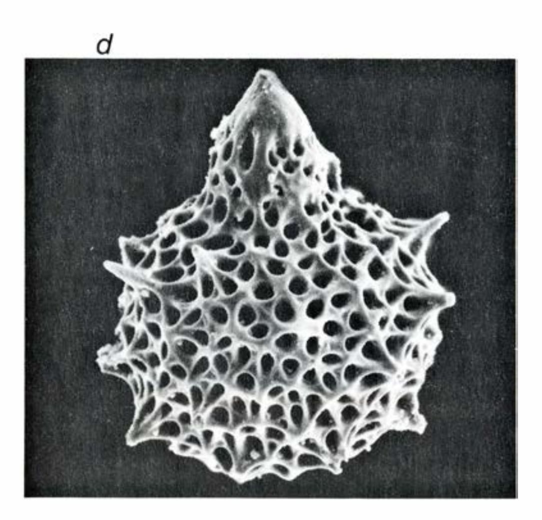
c



2. ESQUELETOS DE RADIOLARIOS, organismos unicelulares que aparecieron por vez primera en los océanos hace unos 500 millones de años. El esqueleto, silíceo, es poco soluble en el agua marina. Hoy pueden separarse de rocas como el pedernal, formado a partir de sedimento oceánico de profundidad. La forma de los esqueletos revela la edad de las rocas, lo que ayuda a establecer la historia de los litosferoclastos. Los dos esqueletos de la izquierda se obtuvieron de argilita del litosferoclasto Golconda, en Nevada septentrional central. Uno de ellos (*a*), del género *Pseudoalbaillella*, tiene unos 290 millones de años de antigüedad; aparece aquí aumentado 200 veces. El otro (*b*), del género *Albaillella*, tiene 250 millones de años; su aumento es de 300 diámetros. Los dos esqueletos de la derecha se extrajeron de pedernal del litosferoclasto de San Simeón, en el sur de California. Uno de ellos (*c*), del género *Archaeodictyomitra*, tiene entre

oceánica en centros expansivos de cresta de dorsal posiblemente se equilibre con la pérdida de corteza oceánica a lo largo de las zonas de subducción. Las pruebas son abrumadoras. El conjunto de rocas corticales producido en centros de expansión, la denominada ofiolita de basalto de dorsal mesoceánica, MORB (por *midocean ridge basalt*), constituida por una sucesión característica de tres estratos que en total alcanzan seis kilómetros de grosor, sólo raramente se observa en cinturones montañosos plegados. (Las cordilleras casi siempre se forman por plegamiento cortical.) Tales cinturones son los lugares donde cabría hallar ofiolitas MORB, si la corteza oceánica se sumara a las masas continentales.

Sin embargo, de la corteza oceánica sobresalen las islas o cadenas insulares creadas por el plácido vulcanismo basáltico propio de los puntos calientes y arcos insulares derivados del vulcanismo más explosivo paralelo a la zona de subducción. Las cuencas oceánicas contienen también fragmentos de márgenes continentales que se separaron de los continentes durante el cuarteamiento de formación de valles de fractura cuando se abrieron océanos nuevos. Además, en el fondo oceánico existen revestimientos de sedimento cuyo volumen actual mundial se cifra en 170 millones de kilómetros cúbicos. Los sedimentos representan acarreos continentales aportados por ríos, restos de fósiles planctónicos y precipitados químicos del mar. Parte del material del fondo oceánico se ha subducido. Con todo, gran parte del sedimento, los montes submarinos basálticos, los arcos insulares volcánicos y los fragmentos continentales están destinados a ser barridos conjuntamente: se convierten en litosferoclastos que incrementan el tamaño de los continentes.



100 y 150 millones de años y está aumentado 700 veces. El otro (*d*), tal vez del género *Stichocapsa*, es de antigüedad y tamaño parecidos. Las fotografías, tomadas con un microscopio de barrido electrónico, las proporcionaron Benita L. Murchey y David L. Jones, del Servicio Geológico de los Estados Unidos.

Naturaleza de los retazos litosféricos

En geología, el término de reciente acuñación "litosferoclasto", o retazo litosférico (equivalente al inglés *terrane* o, en puridad, *tectonostratigraphic terrane*) designa un bloque cortical, de composición no necesariamente uniforme, limitado por fallas. Se trata de una entidad geológica cuya historia difiere de la de los bloques adyacentes. Hay retazos litosféricos, o litosferoclastos, de muchos tamaños y formas, y de grados variados de complejidad en su composición. La India, por ejemplo, es toda ella un gran litosferoclasto. Algunas de sus formaciones rocosas tienen más de 1000 millones de años, y en los últimos 100 millones de años la India se ha comportado como una sola masa. (Fue parte del margen del gran supercontinente Gondwana, hoy descoyuntado, y se rompió y migró al norte hasta chocar con el margen meridional de Asia.) Por el contrario, los litosferoclastos que no se originaron como fragmento de algún continente anterior suelen encerrar una historia bastante sencilla, de menos de 200 millones de años, tiempo máximo habitual de supervivencia de un fondo oceánico. La composición de tales retazos litosféricos tiende a parecerse a la de una isla o meseta oceánica moderna. Algunos litosferoclastos constan principalmente de cantos rodados, arena y limo consolidados; representan abanicos sedimentarios que se acumularon en una cuenca oceánica, por lo general entre fragmentos corticales en colisión.

La geometría de un litosferoclasto es el producto de su historia de movimientos e interacciones tectónicas. Los nacidos en una placa oceánica retienen su forma hasta que chocan y acrecen. Entonces quedan sujetos a movimientos corticales que modifican su forma. Por ejemplo, los de la cordillera de Brooks, en Alaska, son grandes mantos amontonados uno encima de otro. En otras partes de la cordillera de Norteamérica occidental, los litosferoclastos son cuerpos alargados. El alargamiento refleja la rebanadura de la corteza por una malla de fallas noroeste-sudeste, entre las que se cuenta la falla de San Andrés, en California. En Asia, los retazos litosféricos han tendido a conservar las formas que heredaron de episodios de fragmentación y separación; sin embargo, algunos retazos litosféricos menores quedaron aprisio-

nados en choques entre los mayores y se distorsionaron. El conjunto de litosferoclastos de China está estirándose y desplazándose en dirección este-oeste a medida que la India comprime Asia desde el sur.

No siempre se conoce la historia exacta del movimiento de un litosferoclasto dado. En realidad, hace muy poco que se ha logrado documentar las trayectorias de unos pocos retazos. Dado que, por definición, los litosferoclastos están limitados por fallas y difieren de su vecindad geológica, deben haber recorrido, por lo menos, una distancia igual a su dimensión más larga. Las distancias reales varían mucho. Algunos montes submarinos basálticos hoy acrecidos al margen de Oregon se han movido una distancia mínima, a partir de un punto de origen marino situado en las cercanías. En cambio, formaciones rocosas parecidas de alrededor de San Francisco han recorrido hasta 4000 kilómetros a través del Pacífico. Con una velocidad de sólo 10 centímetros anuales, un litosferoclasto errante podría completar un circuito del globo en sólo 400 millones de años. No es de extrañar que los continentes sean confusas aglomeraciones de retazos litosféricos.

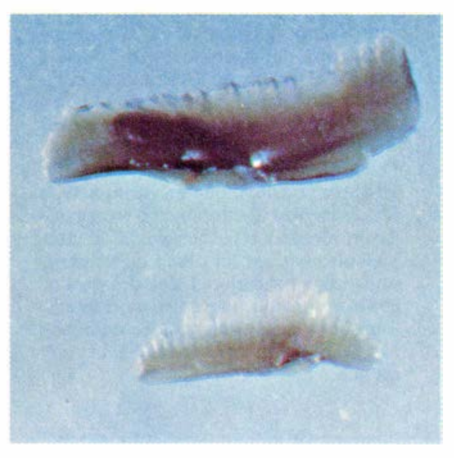
Pruebas

¿Cómo puede reconstruirse la historia de un retazo litosférico? Fundamentalmente, el origen de cada unidad rocosa que compone un litosferoclasto arroja luz sobre la historia evolutiva del conjunto. Las rocas sedimentarias indican ambientes de deposición del pasado: sugieren antiguas gravas fluviales, bajos coralinos, arenas deltaicas, barros de una plataforma continental o barros de un abismo oceánico. También importa su edad, y ayuda a determinarla el registro fósil. Hasta hace poco, sin embargo, el registro resultaba incompleto: la evidencia fósil procedía principalmente de rocas depositadas en antiguos ambientes marinos someros. Sólo a partir de la última década ha podido determinarse la edad de rocas que representan depósitos oceánicos más profundos, rocas que resultan esenciales para comprender los acontecimientos que crearon muchos de los cinturones montañosos de la tierra. En ese avance están implicados radiolarios y conodontos.

Los radiolarios son organismos unicelulares que aparecieron en los océanos ya en el período Cámbrico, hace unos 500 millones de años, y abundaron hasta hace sólo unos 160 millones de años. Ocupaban los niveles superiores del océano, pero su esqueleto era de sílice, sustancia muy poco soluble a cualquier profundidad del mar. En consecuencia, los fangos abisales suelen ser ricos en lo que se llama barro de radiolarios: la acumulación de la sílice de su esqueleto. La roca llamada pedernal, explotada para la confección de puntas de flecha y cuchillos por muchos pueblos, es, de hecho, ese barro endurecido. Perfeccionada la técnica de extracción de radiolarios del pedernal por disolución de la roca en un ácido fuerte, pudieron datarse millares de conjuntos fósiles de pedernal.

Los conodontos son también fósiles microscópicos y, al igual que los radiolarios, hoy pueden extraerse de la roca disolviendo ésta en ácidos fuertes. Costó averiguar su identidad biológica, pero actualmente parece seguro que se trata de restos del esqueleto del aparato alimenticio de un grupo extinguido de pequeños animales marinos parecidos a gusanos, que vivieron entre hace 570 y 200 millones de años. Dado que los conodontos se encuentran asociados a otros conjuntos fósiles, ha podido establecerse una cronología del cambio de morfología de los conodontos. De hecho, para rocas de más de 250 millones de años de antigüedad, la base en la que se apoya la bioestratigrafía por radiolarios (la determinación de la edad de una roca a partir de la naturaleza de los fósiles de radiolarios que contiene) es la coincidencia ocasional de radiolarios y conodontos. Gracias a radiolarios y conodontos pueden hoy datarse muchas rocas cuya edad se desconocía hace sólo una década. A menudo los resultados son sorprendentes. Se encuentran a veces series de rocas viejas apoyadas sobre otras más jóvenes, en lo que parece una mezcla de grandes pilas de estratos. En algunos casos las superficies de apilamiento son paralelas a los estratos y las rocas no muestran indicio obvio de reubicación.

Interviene también la geología estructural en el análisis de los litosferoclastos. La razón es simple: el movimiento de unas rocas frente a otras (o por encima, o por debajo) genera pliegues, aspilleras ("crenelations"), lineaciones y foliaciones, tanto micro como







3. LOS CONODONTOS, a diferencia de los radiolarios, son porciones de organismos mayores: son restos esqueléticos, de no más de un milímetro de longitud, del aparato de alimentación de gusanos marinos extinguidos que vivieron desde hace 570 hasta hace 200 millones de años. Su forma y su ornamentación superficial establecen la edad de la roca en la que se hallan; su color caracteriza la temperatura máxima alcanzada por la roca, cuestión importante en la prospección de hidrocarburos y determinados minerales. (Entre los 50 y 550 grados Celsius la materia orgánica de los conodontos cambia secuencialmente de amarillo pálido a pardo, negro, gris y blanco y, finalmente, pierde todo color; por encima de los 150 grados se pierde toda posibilidad de extraer hidrocarburos líquidos.) Los conodontos de la izquierda, de la cordillera de Brooks, Alaska noroccidental, vivieron en ambientes marinos de agua somera hace unos 360 millones de años. Su color (de amarillo pálido a pardo claro) indica que la roca huésped nunca rebasó los 90 grados Celsius. El conodonto del centro, extraído de un canto rodado pequeño de un conglomerado que aflora a lo largo del río Yukon, en Alaska oriental, vivió a profundidades relativamente someras; el canto es roca que se depositó hace unos 325 millones de años. El color gris indica que la roca alcanzó una temperatura de por lo menos 400 grados. Los conodontos de la derecha, de la región de Glacier Bay, Alaska meridional, vivieron en agua profunda hace aproximadamente 230 millones de años. El color azul-negro indica una temperatura de la roca de por lo menos 300 grados. Las micrografías las proporcionó Anita G. Harris, del Servicio Geológico de los Estados Unidos.

macroscópicas, que ayudan a reconstruir el movimiento. Sin embargo, muchos de los pliegues observados sobre el terreno no representan la dirección primitiva de movimiento, sino otro posterior que consolidó los litosferoclastos en empaquetamientos más trabados. Los datos estructurales se complementan con análisis de la composición de la roca. No parecen existir correlaciones seguras entre asociaciones minerales o composición química y modos de origen particulares, aun cuando caben ciertas generalizaciones. Por ejemplo, el granito cuyo contenido en estroncio es excepcionalmente rico en el isótopo 87 tiende a proceder de la solidificación de magma en un continente antiguo preexistente, mientras que el granito pobre en estroncio 87 indica un origen en un contexto oceánico.

Una contribución importante a la reconstrucción de la historia de los litosferoclastos procede del estudio del paleomagnetismo remanente: el alineamiento de diminutas partículas magnéticas de la roca, inducido por el campo magnético de la tierra en el momento de formarse la roca. En las investigaciones se presupone que el campo magnético terrestre es, en esencia, un dipolo, o barra imantada, coincidente con el eje de rotación del planeta, de modo que las líneas de campo en el ecuador son horizontales (paralelas a la superficie de la tierra) y las líneas de campo en cada polo son verticales (perpendiculares a la superficie). Entre el ecuador y los polos las líneas de campo se van empinando. La mayoría de las rocas sedimentarias y volcánicas se depositan horizontalmente, en capas; por consiguiente, la inclinación de su paleomagnetismo remanente puede revelar la latitud a la cual se formó la roca. Además, la dirección geográfica de la orientación del magnetismo sugiere el giro que, en algún momento de sus migraciones geológicas, pueda haber dado la roca.



4. UN VASTO OCEANO, el Panthalassa, dominaba la superficie terrestre hace 250 millones de años, cuando la práctica totalidad de la corteza continental del planeta se aglomeraba en un solo supercontinente, Pangea. Desde entonces, Pangea se ha fragmentado en los actuales continentes y la corteza oceánica de la cuenca del Panthalassa se ha subducido íntegramente (ha retornado de nuevo al manto terrestre). Su puesto lo ocupan las cuencas oceánicas actuales. Los restos de Panthalassa (constituidos por fragmentos de corteza continental junto con corteza creada por actividad volcánica) produjeron litosferoclastos que incrementan los continentes alrededor del Pacífico. Tales litosferoclastos aparecen en naranja en las figuras 5 y 6.

Tectónica de mosaico

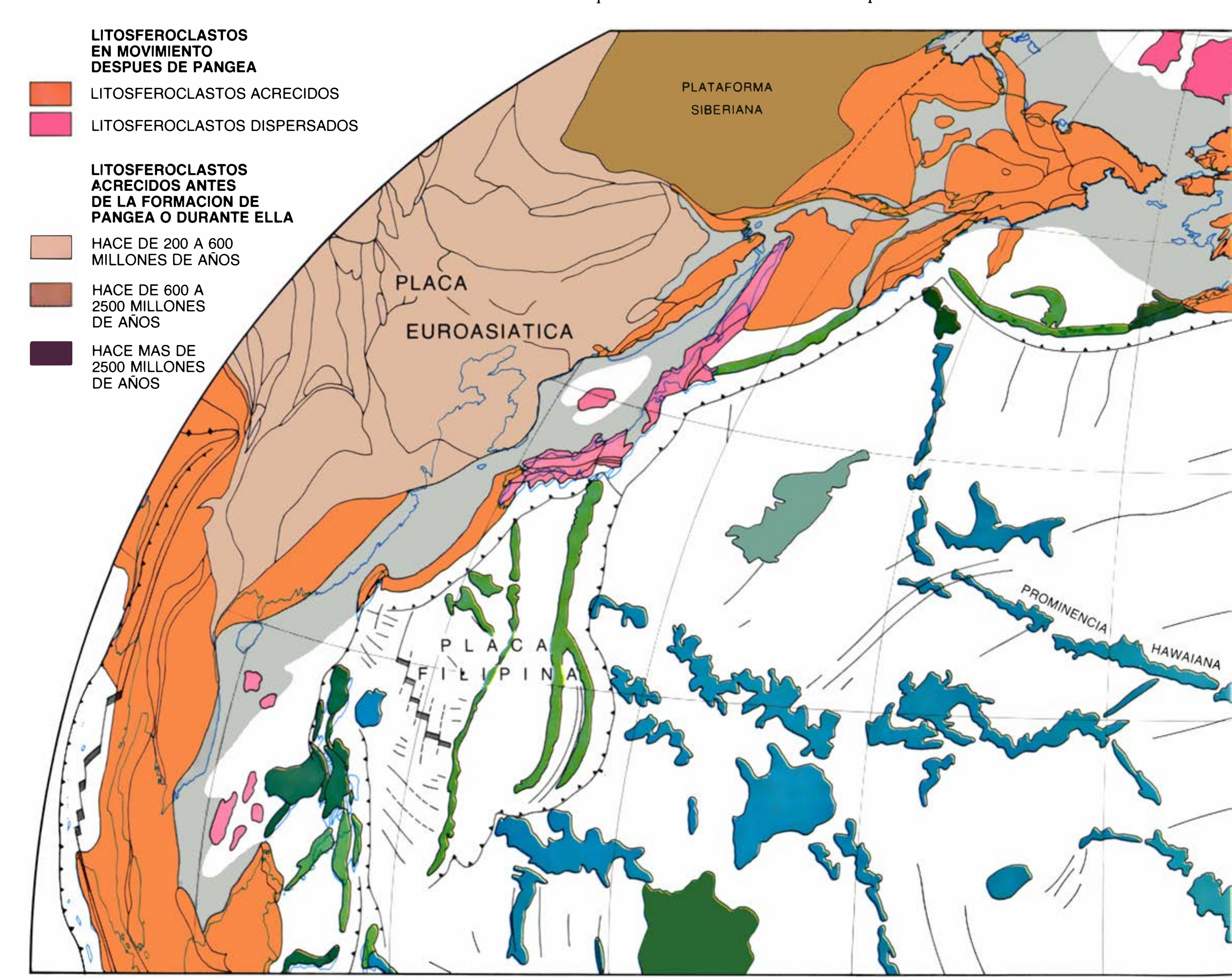
A partir de todas esas líneas de argumentación mostrativa abordaré, en términos de litosferoclastos, la reconstitución y el remodelamiento de los continentes, en el pasado y en el futuro. Vale la pena recordar algunas cifras. El volumen planetario actual de la corteza continental es de aproximadamente 7600 millones de kilómetros cúbicos; la roca más antigua conocida tiene 3800 millones de años. Dividiendo la primera cifra por la segunda se obtiene una estimación lineal simple de que los continentes han crecido a un ritmo mundial de unos dos kilómetros cúbicos anuales, unos 65 metros cúbicos por segundo. El cálculo quizá peque de elevado; los procesos de crecimiento en la caldeada tierra primitiva tal vez fueran más rápidos que el promedio. Se han propuesto muchas curvas de crecimiento; la mayoría presuponen que del 70 al 80 por ciento de todo el crecimiento cortical se produjo hace más de 2000 millones de años. El 20 o 30 por ciento final de la actual masa de los continentes se habría acumulado durante los últimos 2000 millones de años, a un ritmo medio de entre 0,7 y 1,1 kilómetros cúbicos anuales, tasa muy en consonancia con la de las contribuciones a la corteza que aportan los arcos volcánicos y montes submarinos oceánicos modernos.

La acumulación equivale al emplastamiento de litosferoclastos en cratones (núcleos continentales) preexistentes, las porciones más antiguas de la corteza continental. El proceso puede seguirse con máximo detalle a través del Fanerozoico, el intervalo de aproximadamente 600 millones de años del que se dispone de un abundante registro fósil de vida pluricelular. A principios de ese intervalo, los continentes (según los datos paleomagnéticos) eran masas aisladas desparramadas alrededor del globo por la región ecuatorial. (El período comprendido entre hace 700 y 500 millones de años fue, aparentemente, una época de gran fragmentación continental.) En los 350 millones de años siguientes, el desplazamiento de los continentes se tradujo, primero, en la aglomeración de dos megacontinentes, Gondwana y Laurasia, y, luego, hace 250 millones de años, en la unión de ambos en el supercontinente Pangea, una masa que, a grandes rasgos, tenía forma de cuarto creciente y una orientación general norte-sur. Los viejos núcleos continentales, aumentados por la acumulación de litosferoclastos desde principios del Fanerozoico, empezaron de nuevo a fragmentarse, hace unos 200 millones de años, a lo largo de un nuevo patrón de valles de fractura, o rifts, parecido al de los

actuales 56.000 km de centros de expansión oceánica.

Imaginemos, dentro de varios miles de millones de años, la formación de un supercontinente nuevo constituido por Asia y América del Norte y del Sur. El Pacífico se habrá cerrado, tras la subducción de la dorsal expansiva del Pacífico oriental, mientras el Atlántico habrá proseguido su ensanchamiento. También puede predecirse que para entonces los continentes en colisión habrán aumentado de tamaño. La actual extensión superficial de los continentes que rodean el Pacífico alcanza los 290 millones de kilómetros cuadrados, de los cuales los litosferoclastos de nueva acreción (es decir, el material posterior al de Pangea) han aportado aproximadamente 25 millones de kilómetros cuadrados, el nueve por ciento. Si se supone que la corteza tiene un grosor medio de 20 kilómetros, el ritmo de crecimiento cortical en el Pacífico durante los últimos 200 millones de años ha sido de 2,5 kilómetros cúbicos por año. La cifra resulta algo engañosa: entre los mosaicos tectónicos que orlan el Pacífico se cuentan algunos litosferoclastos grandes constituidos por corteza continental desplazada, formada antes de la rotura de Pangea. Entre los ejemplos se cuentan la mitad oriental de México, la cordillera de Brooks, en Alaska, porciones de la Unión Soviética nororiental y la mayor parte de la península Malaya. Con todo, las investigaciones preliminares parecen indicar que los ritmos de crecimiento de la corteza continental que circunda el Pacífico superaron el ritmo medio planetario de un kilómetro cúbico anual.

Resultan interesantes al respecto algunos estudios recientes de paleomagnetismo remanente en bloques calcáreos de California septentrional. Los estudios indican que la caliza, de 85 a 100 millones de años de antigüedad, se depositó al sur del ecuador. Y sin embargo, la edad de la roca sedimentaria que hoy se solapa con los bloques sugiere que la caliza (y los basaltos asociados) se anexionaron al margen de California no después de hace 38 millones de años. Los datos exigen que la placa que acarreaba la caliza se haya movido hacia el norte a una velocidad de entre 15 y 30 centímetros por año, más deprisa que el movimiento actual de las placas. Esos valores me incitan a



5. LITOSFEROCLASTOS DEL PACIFICO SEPTENTRIONAL. Dominan el mapa correspondiente a algo más de una cuarta parte de la superficie terrestre. El centro expansivo del Pacífico se halla muy al este del centro del Pacífico; aquí se ve parte del mismo al sur de América Central. El fondo del Pacífico se expande desde ese punto, para hundirse a lo largo de grandes fosas, que señalan las zonas de subducción ubicadas a lo largo de los márgenes del Pacífico. Los colores indican las edades en que acrecieron los litosferoclastos. Los primeros (*pardos*), de más de 2500 millones de años, comprenden los antiguos núcleos de los continentes. Las acreciones más jóvenes (*naranja*), que corresponden al nueve por ciento del área superficial de los continentes que circundan el Pacífico, consti-

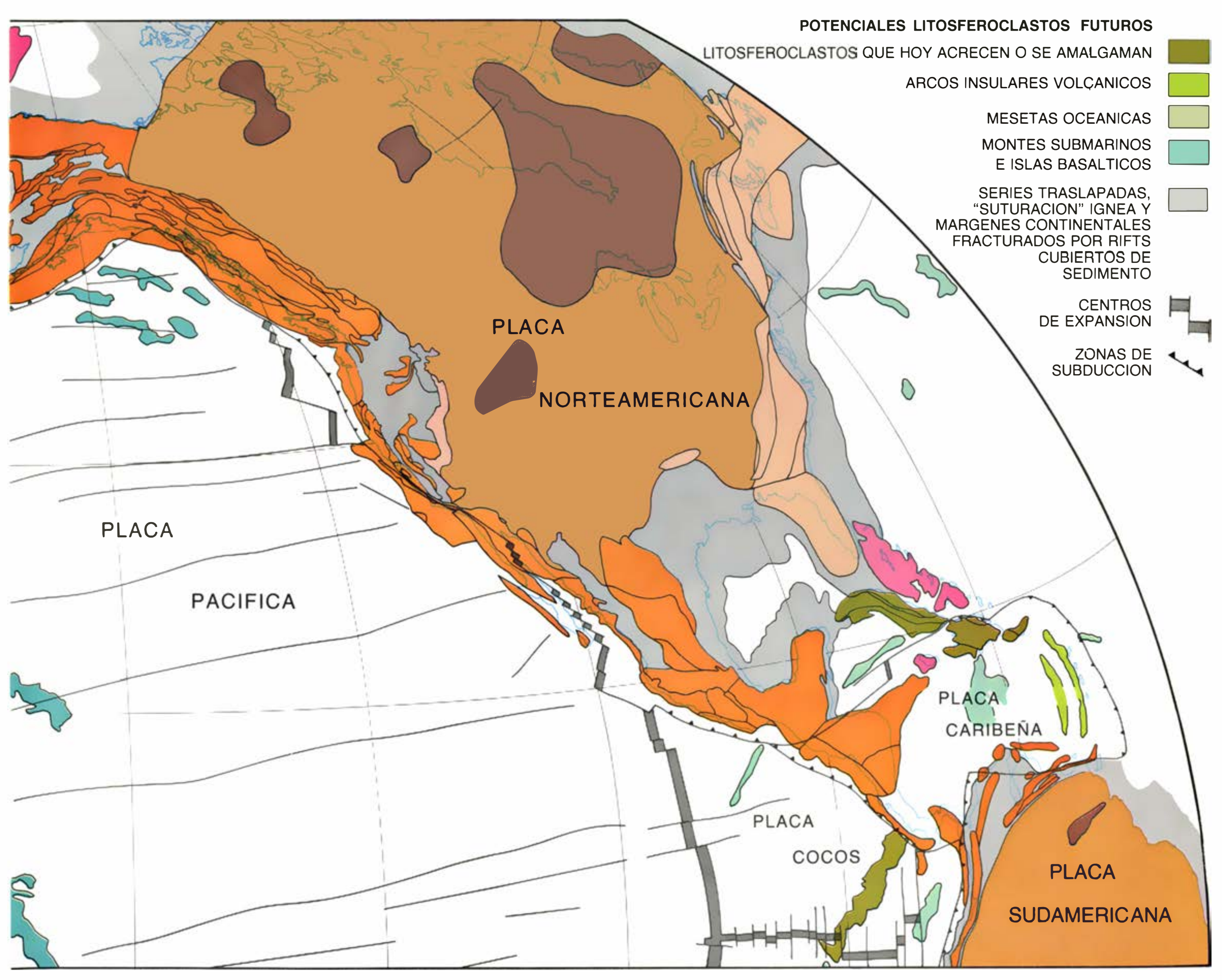
especular que el ritmo de crecimiento de los continentes también es variable. Tal vez siga ciclos de centenares de millones de años de duración.

La importancia de los sedimentos

En el incremento o la reducción de los continentes, el papel de los sedimentos no es, ni mucho menos, pasivo; también pasan a formar parte de litosferoclastos. Por una parte, gruesas pilas de sedimentos permanecen en los continentes y se acumulan a lo largo de sus fracturados márgenes. Sólo un 30 por ciento del sedimento depositado por ríos rebasa el margen continental y se decanta en corteza oceánica. Por otro lado, parte de los sedimentos trasladados a zonas de subducción por la cinta transportadora oceánica parece empotrarse frente a la placa cortical cabalgante en forma de "prisma de acreción", o queda adherida a la cara inferior de la placa cabalgante. También en ese caso suelen quedar atrapadas grandes acumulaciones de sedimentos entre masas corticales que chocan. Puede observarse un ejemplo actual en el mar de las Molucas, donde chocan dos arcos insulares. En otras ocasiones la propia pila de sedimentos constituye un litosferoclasto; tal es el caso de los prismas de acreción en la región de la isla Kodiak y el golfo de Alaska. O, por el contrario, la pila puede constituir la matriz donde se engloben litosferoclastos, como en la cordillera de Alaska.

En el mundo actual la mayor fuente aislada de sedimento es el encumbrado accidente del relieve resultante del choque de la India y Asia. La corteza asiática ha cobijado el litosferoclasto indio, duplicando el grosor de la corteza y levantando las montañas del Himalaya y, a su norte, la meseta tibetana. Seis grandes sistemas fluviales, el río Amarillo, el Changjiang, el Irauadi, el Mekong, el Ganges-Brahmaputra y el Indo, drenan la región, que constituye sólo el cuatro por ciento de la superficie de terreno del mundo. Juntos descargan en los océanos unos 3800 millones de toneladas de sedimento por año, casi el 40 por ciento del sedimento descargado por todos los ríos de la tierra.

El sedimento está compuesto por limo y arcilla, junto con roca y granos minerales. También hay agua que, debido a la porosidad de los sólidos, re-



tuyen residuos corticales barridos por la subducción del Panthalassa. Las mesetas oceánicas, los montes submarinos y los arcos insulares modernos representan posibles litosferoclastos futuros. En el lado norteamericano del Pacífico, la costa aparece compuesta por retazos litosféricos procedentes de restos oceánicos, incluidos arcos insulares dilatados en astillas. El norte de Alaska (la cordillera de Brooks), lo forman litosferoclastos cabalgados unos sobre otros y compuestos de corteza de margen continental y oceánica. En el lado asiático del Pacífico, la corteza continental contiene litosferoclastos antiguos rodeados de cinturones de retazos litosféricos más modernos. El viejo núcleo de Asia constituyó el tope trasero a partir del cual se acumularon litosferoclastos más jóvenes.

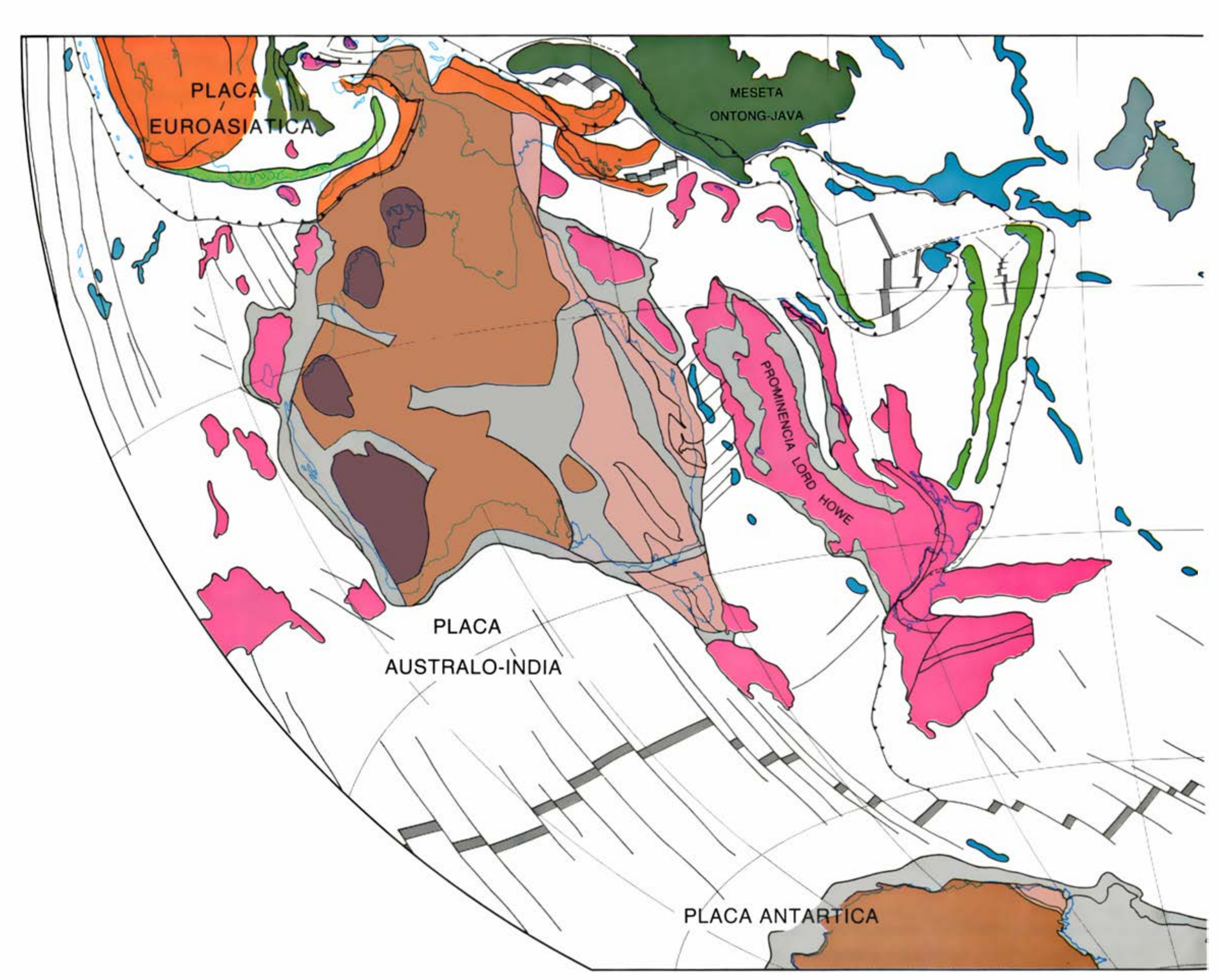
presenta a menudo hasta el 50 por ciento. Si la densidad neta del sedimento es de dos gramos por centímetro cúbico, el volumen del sedimento descargado anualmente desde los ríos asiáticos es de 1,7 kilómetros cúbicos: el volumen mundial se sitúa entre 4,5 y 6,8. A medida que el sedimento se compacta y petrifica, la porosidad disminuye hasta casi desaparecer. Por consiguiente, la descarga mundial aporta de 3,3 a 4,9 kilómetros cúbicos de roca por año. (He presupuesto una densidad de la roca de 2,75 gramos por centímetro cúbico, valor algo superior a la densidad del cuarzo común.)

Se desconoce el destino a largo plazo de la mayor parte de la roca. Cierta proporción puede subducirse; otra parte quizá se levante entre macizos continentales que chocan; otra podría verse arrancada de su lugar de acumulación (por ejemplo de un delta fluvial o de un depósito oceánico profundo) para terminar acreciendo un continente distante. Con todo, los 3,3 o 4,9 kilómetros cúbicos de roca que pueden formarse cada año por sedimentación a lo largo de márgenes continentales, y sobre el fondo oceánico, rebasan ampliamente los 0,2 kilómetros cúbicos que aporta anualmente el vulcanismo basáltico y los 0,75 a 1,5 kilómetros cúbicos procedentes del vulcanismo explosivo de zona de subducción. No cabe duda de que la roca sedimentaria, o su equivalente metamorfizado, es un componente principal de los cinturones montañosos plegados. Ciertamente, los cálculos que he indicado sugieren que hasta el 75 por ciento de corteza continental de nueva formación podría consistir en sedimento y productos metamorfoseados o fundidos de sedimento.

Retazos litosféricos del Pacífico septentrional

La mejor manera de investigar la construcción y el modelado de continentes a partir de litosferoclastos es examinar regiones específicas del planeta. Me intereso por el Pacífico, al que divido en cuadrantes. Cada cuadrante muestra un patrón geológico característicamente diferenciado, reflejo de historias contrastadas de acreción y dispersión de litosferoclastos.

En el cuadrante nororiental, donde el Pacífico baña la costa de Norteamérica, se han amontonado durante los últimos 180 millones de años litosfero-

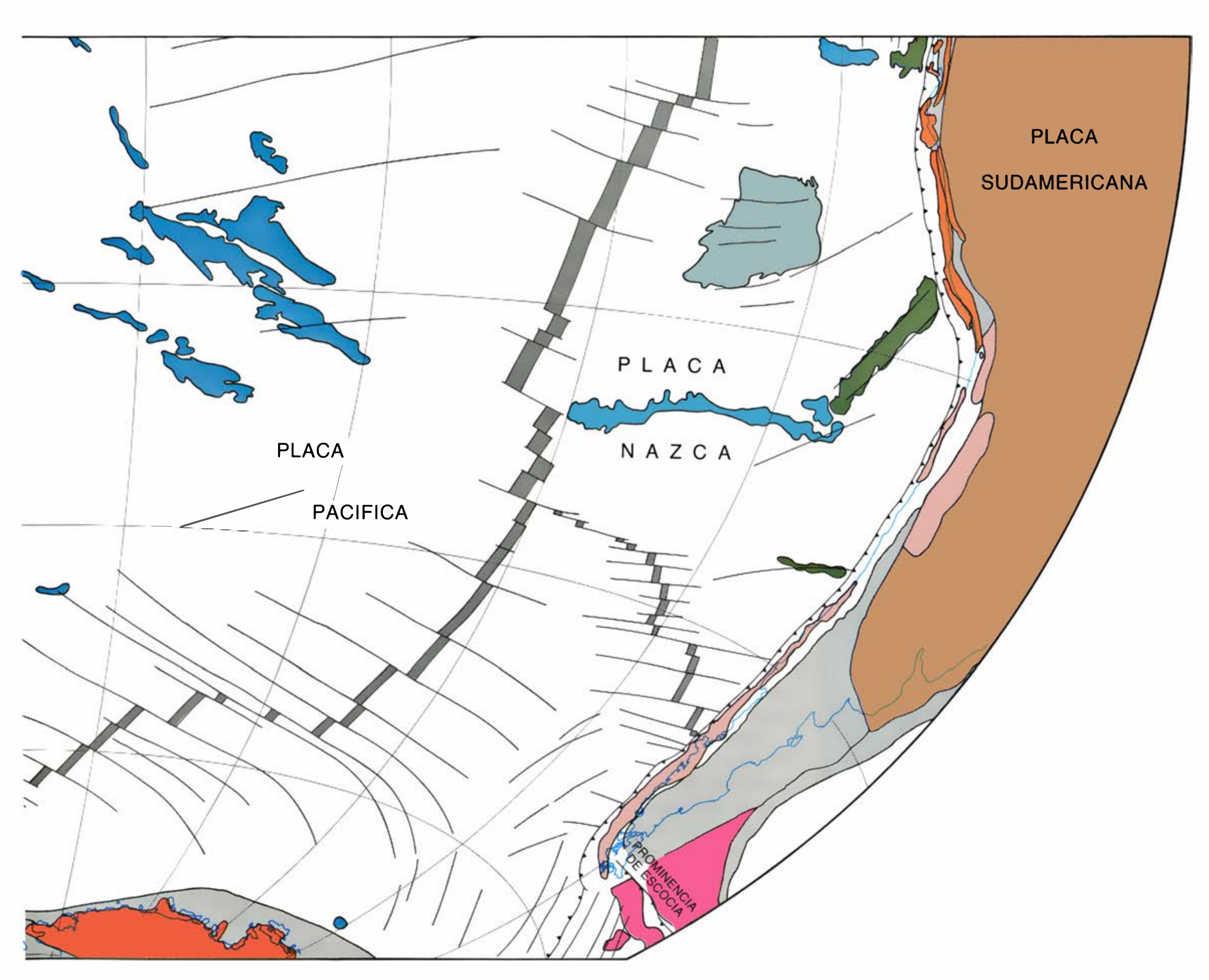


6. LITOSFEROCLASTOS DEL PACIFICO AUSTRAL. Ocupan un mapa que, complementando la ilustración anterior, muestra el Pacífico entero. En la parte occidental del mapa, la Antártida, Australia y Nueva Zelanda señalan los restos del valle de fractura (rift) de tres brazos que empezó a abrirse hace de 120 a 100 millones de años en la parte austral de Pangea, que se formó a partir del megacontinente Gondwana. Sigue activa la acreción de litosferoclastos al noreste de Australia. La meseta oceánica Ontong-Java (un fragmento de corteza continental) ya se ha incrementado a lo largo de su borde meridional con parte

clastos derivados principalmente de arcos insulares y otros materiales oceánicos. La cordillera costera comprende una malla de fallas de desgarre orientadas de oeste a este. El movimiento es aquí predominantemente horizontal, a lo largo del plano de la superficie terrestre; los litosferoclastos de nueva acreción se han dilatado en una serie de astillas. Wrangellia, un litosferoclasto bien estudiado que un día se halló en el ecuador, o incluso más al sur, constituye un buen ejemplo. Wrangellia se engarzó en Oregon hace unos 70 millones de años. El ulterior fallamiento ha esparcido fragmentos de Wrangellia hacia el norte, dejando partes en Oregon oriental, en las islas Vancouver y Reina Carlota y por las montañas Wrangell de Alaska meridional. Las partes central y occidental de Oregon son litosferoclastos barridos después del emplazamiento de Wrangellia.

La cordillera de Brooks muestra una historia bien diferente. Mantos finos y extensos de estratos, representativos de un margen continental, han cabalgado unos sobre otros, transformando una geografía que un día midiera 500 por 1000 kilómetros, por lo menos, en un montón cortical de 500 por 300 kilómetros. El análisis de indicios sobre las direcciones de flujo de fluidos cargados con sedimento que contribuyeron a los estratos indica que el montón entero debe haberse desplazado hasta su ubicación actual, pero aún se discute su procedencia. Una hipótesis postula que la pila emergió de las islas del Artico canadiense mediante un giro de sentido contrario al de las manecillas del reloj. La cuenca canadiense del océano Artico representaría, pues, la depresión dejada por la masa continental giratoria.

El cuadrante noroccidental del Pacífico incluye Asia, Japón y las Filipinas. Aquí la corteza continental consta de viejos fragmentos continentales, cada uno de ellos flanqueado, o incluso rodeado, por cinturones de litosferoclastos acrecidos durante la era Paleozoica, hace entre 600 y 250 millones de años. En efecto, la plataforma siberiana parece haber constituido el tope trasero a partir del cual se emplazó hacia el exterior una sucesión de litosferoclastos. A lo largo del límite meridional de la plataforma, arcos volcánicos y otros trozos corticales se amontonaron a principios del Paleozoico, entre hace 600 y 400 millones de años; formaron el cinturón plegado del Baikal,



del arco insular de las Nuevas Hébridas. Pueden unírsele otros arcos, y quizá acabe todo el conjunto anexionándose a Australia. En la parte oriental del mapa, la placa pacífica se subduce bajo la placa sudamericana. Al sur de la punta de Sudamérica se forman litosferoclastos nuevos: la corteza oceánica del mar de Escocia se abre hacia el este entre Sudamérica y la península Antártica, tallando litosferoclastos a lo largo de sus márgenes antes de subducirse. Esta y la figura anterior muestran los resultados de los investigaciones de David L. Jones, Erwing Scheibner, Zvi Ben-Avraham, Elizabeth R. Schermer y del autor.

la región montañosa situada entre Mongolia y el mar de Okhost. Luego, entre hace unos 300 a 60 millones de años, cuando se empotró la India, varios litosferoclastos (Tarim, Chiangjiang, el macizo chino-coreano, Indochina y, finalmente, la India) se juntaron formando Asia.

Hoy asistimos a un episodio de acreción parecido. La meseta oceánica Ontong-Java probablemente sea un fragmento fracturado de corteza continental. En la actualidad se halla sumergida en su mayor parte. Su tamaño es comparable al del litosferoclasto asiático de Chiangjiang. Acrecida junto al lado meridional de la meseta Ontong-Java, queda una porción del arco volcánico de Nuevas Hébridas. En los aledaños hay muchos otros arcos. Si el mar del Coral se cierra –acontecimiento difícil de pronosticar, por cuanto en algunos puntos el mar se abre, mientras que en otros se cierra– la meseta de Ontong-Java, orlada de arcos volcánicos acrecidos, posiblemente se convierta en un importante añadido a Australia.

Retazos litosféricos del Pacífico austral

El cuadrante sudoccidental del Pacífico, que incluye la Antártida, Australia y Nueva Zelanda, es una región caracterizada tectónicamente por una dispersión radial: las masas continentales son el producto de la rotura de parte de Gondwana, iniciada entre hace 120 y 100 millones de años, cuando se desarrolló un sistema de valles de fractura de tres brazos. Un brazo creó el mar de Tasmania; los otros dos separaron la Antártida de Australia y la meseta de Campbell de Nueva Zelanda. Los cinturones plegados de la Antártida occidental, Australia oriental y Nueva Zelanda sugieren la antigua historia de la región. Evidentemente, por episodios de acreción se edificaría la corteza continental, hacia fuera, a partir de núcleos hoy emplazados en la Antártida oriental y Australia occidental.

Queda el cuadrante sudoriental del Pacífico, que comprende la costa occidental de Sudamérica. Lo forman regiones muy variadas. En la parte austral del cuadrante se están tallando litosferoclastos nuevos en el mar de Escocia, el corte entre los Andes australes y la península Antártica, y quedan pequeños bloques corticales limitados por fallas, como las islas Georgia del Sur y Orcadas. Hacia el norte, desde el sur de Chile hasta Perú, los Andes se extienden en línea casi recta a lo largo de 3000 kilómetros. La región reclama una investigación más profunda pero, por lo que hoy se sabe, apenas parece presentar signos de acreción de litosferoclastos, a pesar de los últimos 200 millones de años de subducción. (Se está subduciendo corteza oceánica bajo la corteza continental de Chile y Perú.)

Desde el centro del Perú hasta el Caribe, la parte occidental de Sudamérica muestra una textura de retazos, de litosferoclastos acrecidos. Aquí, como en el extremo austral de Sudamérica, sigue activa la dispersión cortical. Por ejemplo, los sedimentos petrolíferos del lago de Maracaibo, en el noroccidente de Venezuela, rellenan parte de una comba de la corteza dejada cuando el arco insular antillano de la placa del Caribe se deslizó al este, resbalando frente al borde septentrional de Sudamérica. Como dos motoniveladoras, el arco de Escocia, en el sur, y el arco de las Antillas, en el norte, avanzan hacia el Atlántico, quedando a lo largo de los flancos del avance cascotes corticales, que previsiblemente constituirán materia prima para la construcción de futuros litosferoclastos.

El concepto de retazo litosférico, o litosferoclasto, se incorpora a la teoría de la tectónica de placas a medida que se van examinando patrones y procesos de crecimiento continental. El balance global de crecimiento continental es dinámico. Los volcanes aportan entre 0,75 y 1,5 kilómetros cúbicos al año, mientras que los puntos calientes contribuyen con 0,2 kilómetros cúbicos anuales. Simultáneamente, por erosión, los continentes pierden al año hasta 4 kilómetros cúbicos, si bien se recicla hasta tres cuartas partes de lo perdido: el sedimento fruto de la erosión se levanta, se pliega, a veces se metamorfosea y, en ocasiones, se funde en los procesos acrecionarios, de choque, que levantan montañas a lo largo de los márgenes continentales. Al promediar miles de millones de años de historia geológica probablemente se enmascaren períodos especialmente activos, o incluso ciclos de crecimiento continental. Producto de la rotura de Pangea, hace 200 millones de años, se consumió todo un océano de dimensiones planetarias, el Panthalassa, creciendo la corteza continental circumpacífica a un ritmo de hasta 2,5 kilómetros al año, superior, por lo que se ve, al ritmo medio de crecimiento a largo plazo. Por consiguiente, los acontecimientos del Pacífico parecen exigir períodos complementarios de tranquilidad tectónica. El reto actual de la geología es la cartografía de la historia detallada de la tierra a través de los litosferoclastos.

Respiración cutánea en los vertebrados

Puede complementar o reemplazar a la respiración mediante pulmones o branquias. Adaptaciones especiales de la piel y del sistema circulatorio contribuyen a la regulación del intercambio cutáneo de oxígeno y CO_2

Martin E. Feder y Warren W. Burggren

Debido, quizás, a que el hombre respira casi exclusivamente mediante sus pulmones, existe la creencia generalizada de que la respiración tiene lugar sólo en los órganos especializados: si no en los pulmones, en las branquias de los peces y de los crustáceos, en las tráqueas de los insectos o en los pulmones laminares de las arañas. No obstante, siempre que una membrana relativamente fina separa el medio respiratorio (el aire o el agua que un animal respira) de las células vivas o de la sangre circulante, el oxígeno puede penetrar en las células y el anhídrido carbónico puede eliminarse de ellas. Datos recientes que examinaremos aquí sugieren que la piel se utiliza como órgano de intercambio gaseoso eficaz y muy bien regulado en muchos vertebrados.

El intercambio gaseoso cutáneo ha intrigado durante mucho tiempo a los fisiólogos. August Krogh abordó ya la cuestión a principios de siglo en Dinamarca. Krogh obstruyó el flujo de aire hacia el pulmón en ranas y observó que, durante el invierno, cuando las ranas permanecen normalmente inactivas, la piel podía aportar suficiente oxígeno a la sangre; sin embargo, durante las otras estaciones se demostró que los pulmones son necesarios. Las investigaciones de Krogh sobre la circulación de la sangre en los capilares le valieron en 1920 la consecución del premio Nobel en fisiología o medicina.

Durante los años sesenta y setenta, muchos estudios experimentales analizaron la distribución del intercambio gaseoso entre los diferentes órganos respiratorios de un animal: la piel, los pulmones y las branquias, cuando existen. En ese contexto, Victor H. Hutchinson y sus alumnos, de la Universidad de Rhode Island, colocando máscaras faciales de plástico a salamandras, determinaron qué proporción del total del oxígeno utilizado por el animal se captaba a través de la piel y qué proporción del total del anhídrido carbónico se eliminaba a través de ella. Los resultados de estos y de muchos otros estudios similares dieron lugar a una lista sorprendentemente amplia de vertebrados que respiran a través de la piel.

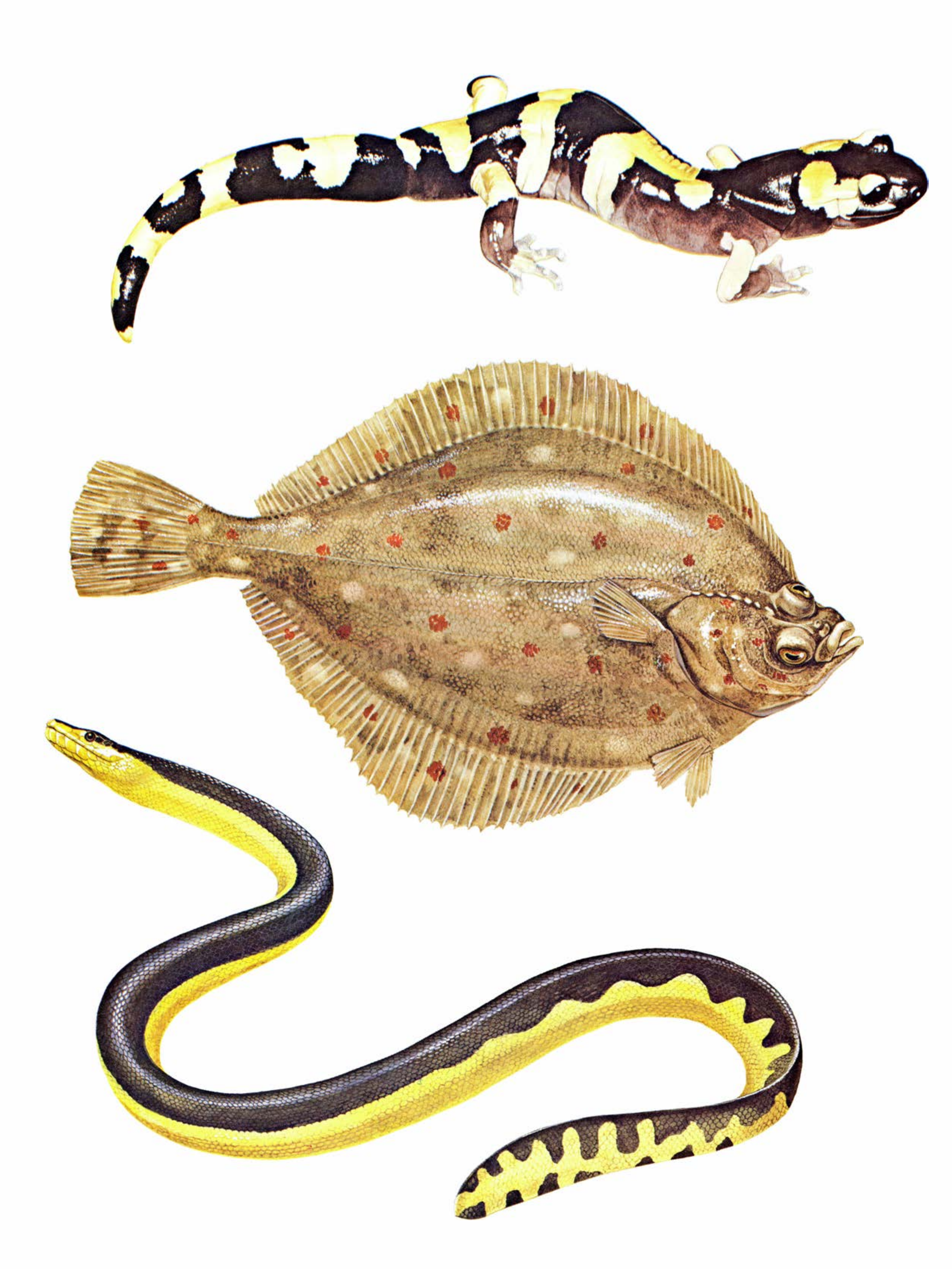
Los animales con respiración cutánea mejor estudiados son probablemente los anfibios. Entre éstos, no es raro que al menos el 30 por ciento de captación total de oxígeno, e incluso el 100 por ciento de la eliminación de anhídrido carbónico, se produzca a través de la piel. Las larvas de rana, por ejemplo, intercambian aproximadamente el 60 por ciento de sus gases respiratorios a través de la piel, aun disponiendo tanto de branquias como de pulmones.

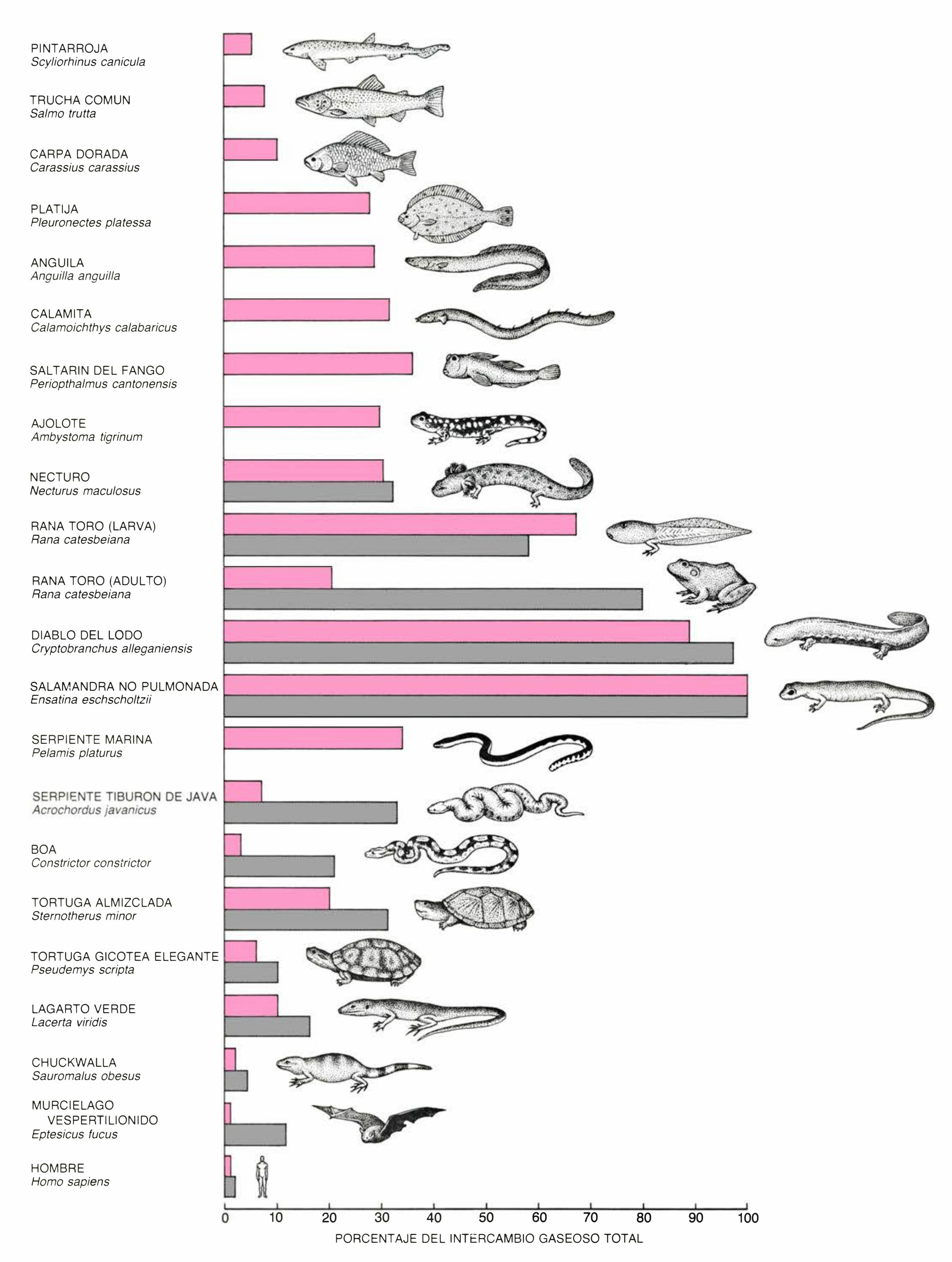
Los anfibios que permanecen casi todo el tiempo en tierra, no en el agua, se enfrentan con dificultades que ponen en un compromiso potencial su supervivencia, porque las mismas propiedades que hacen de la piel una membrana eficaz para el intercambio gaseoso favorecen también la pérdida de agua. Aun así, en todas las especies de anfibios terrestres que han sido investigadas al respecto se ha detectado intercambio gaseoso cutáneo. De hecho, la piel es el único órgano respiratorio en las salamandras adultas de la familia Plethodontidae. Veintenas de especies de esta familia se encuentran en ambientes terrestres tan diversos como los sotobosques de Nueva Inglaterra y las bóvedas de vegetación de las selvas tropicales. Aunque algunos pletodóntidos tropicales alcanzan pesos corporales de hasta 150 gramos y longitudes de 24 centímetros, llevan a cabo todo el intercambio gaseoso entre los tejidos y el medio ambiente a través de la piel.

La respiración cutánea ha sido también objeto de investigación en otros vertebrados, desde los peces hasta los mamíferos. La experimentación ha demostrado, por ejemplo, que los peces con respiración aérea utilizan su piel para casi la mitad de su intercambio gaseoso, en particular si se aventuran a permanecer fuera del agua: las branquias característicamente se colapsan en el aire y pierden su utilidad. Muchos biólogos, que dan por supuesto que los pulmones constituyeron un prerrequisito para la evolución de los animales terrestres, no han prestado atención a la adaptación de la piel como órgano para la respiración aérea.

Peces de los grupos más comunes utilizan también el intercambio gaseoso cutáneo. Los tiburones, el bacalao, la trucha y la carpa dorada –por mencionar algunos– obtienen entre el 5 y el 30 por ciento del total del oxígeno que utilizan a través de la piel. Aparentemente, ni la estructura del tegumento

1. TRES VERTEBRADOS CON RESPIRACION CUTANEA, ilustrados en la página opuesta: una salamandra no pulmonada, una platija y una serpiente marina. La salamandra pertenece a la especie *Ensatina eschscholtzii* y vive en los bosques de California y de Oregón; mide unos cinco centímetros de largo cuando alcanza su completo desarrollo. La piel es el único órgano respiratorio en los adultos de esta especie. La platija *Pleuronectes platessa*, pez plano del Atlántico Norte con valor comercial, capta el 27 por ciento de su oxígeno a través de la piel. El pez puede llegar a alcanzar un peso de cinco kilogramos y una longitud de 90 centímetros. La serpiente marina *Pelamis platurus* absorbe más de un tercio de su oxígeno y elimina más de las tres cuartas partes de su anhídrido carbónico metabólico por vía cutánea. La serpiente alcanza aproximadamente un metro; vive en la zona tropical del Océano Indo-Pacífico, entre América y Africa.

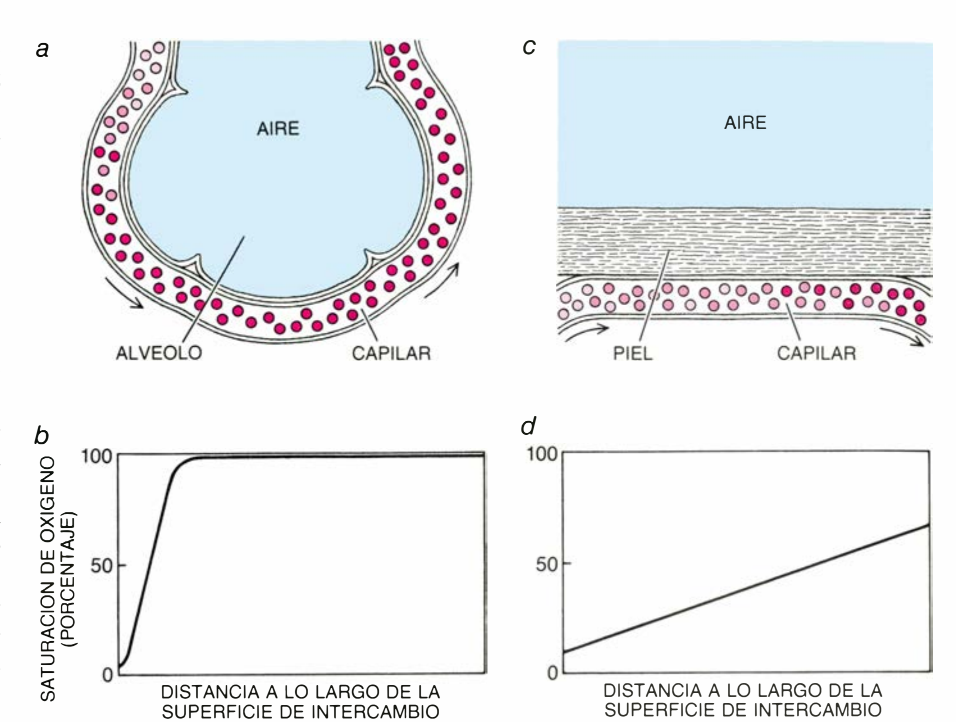




2. EL INTERCAMBIO GASEOSO CUTANEO está muy extendido entre los vertebrados. Aunque predomina más en los anfibios, este tipo de respiración tiene también importancia en otros grupos. La eliminación cutánea de anhídrido carbónico (*barras grises*) constituye de forma característica una fracción del total del intercambio gaseoso mayor que la captación de oxígeno (*barras de color*) en los casos en que se han realizado determinaciones sobre ambos gases.

ni la forma corporal determinan la proporción en que un determinado pez respira mediante su piel. Tanto las platijas, planas y sin escamas, como los calamitas, peces alargados y densamente escamosos, captan aproximadamente una tercera parte del oxígeno que necesitan por vía cutánea.

También se ha demostrado que los reptiles se benefician del intercambio gaseoso cutáneo. Algunas tortugas de agua dulce, a pesar de sus gruesos caparazones, dependen en gran proporción, o incluso totalmente, del intercambio gaseoso cutáneo, sobre todo si pasan el invierno en charcas heladas. Tanto las serpientes que viven en agua dulce como las marinas sacan partido de su piel como órgano respiratorio. Estos reptiles suelen sumergirse durante largos períodos de tiempo y complementan los depósitos de oxígeno del pulmón y de la sangre con oxígeno obtenido a través de la piel. Roger S. Seymour, de la Universidad de Adelaida, en Australia, ha sugerido incluso que algunas serpientes marinas utilizan su piel para eliminar nitrógeno en inmersiones prolongadas, previniendo en consecuencia el síndrome de la descompresión (formación de burbujas de nitrógeno en la sangre) cuando emergen. Muchos reptiles que viven en ambientes áridos, donde la deshidratación representa un peligro constante, han contrarrestado este problema desarrollando gruesas escamas, un caparazón o pieles correosas. Algunas especies, como el lagarto chuckwalla (*Sauromalus obesus*) de los desiertos del sudoeste norteamericano, desarrollan algún intercambio gaseoso cutáneo a pesar de sus gruesas pieles protectoras.



3. CAPTACION CUTANEA DE OXIGENO. Está limitada por el espesor de la barrera de difusión de la piel: la distancia que existe entre el medio respiratorio (el aire o el agua) y la sangre. El oxígeno puede difundirse rápidamente a través de la fina membrana alveolar del pulmón de un mamífero (*a*); la hemoglobina de los glóbulos rojos se oxigena pronto (*rojo oscuro*) y este gas satura completamente la sangre de los capilares adyacentes (*b*). Debido a que la difusión del oxígeno es mucho más lenta a través de la piel, en razón de su mayor espesor (*c*), la sangre de los capilares cutáneos nunca llega a saturarse totalmente de oxígeno (*d*).

Aunque la respiración cutánea es con frecuencia significativa e incluso crucial en los vertebrados inferiores, la piel rara vez es una vía importante para el intercambio gaseoso entre los vertebrados superiores, como las aves y los mamíferos. Incluso en el caso de que las pieles peludas o plumosas de estos animales fueran tan permeables como la piel de las ranas, los vertebrados superiores necesitarían de sus pulmones para mantener su muy superior tasa metabólica. Los pulmones aportan una superficie para el intercambio de gases más amplia y delgada que la que la piel pueda representar. Sin embargo, esta generalización presenta excepciones. Clyde F. Herreid II y sus colaboradores, de la Universidad de Duke, encontraron que los murciélagos eliminaban hasta un 12 por ciento de su producción total de anhídrido carbónico a través de las delgadas y bien vascularizadas membranas alares.

Los mamíferos y las aves nacen o eclosionan con pieles habitualmente finas, densamente vascularizadas y que carecen de pelaje o plumas. En consecuencia, la respiración cutánea puede adquirir en las formas en desarrollo de estos vertebrados superiores una importancia mayor que en los adultos. Con toda probabilidad, en los embriones y fetos de los mamíferos se da un grado de intercambio gaseoso cutáneo significativo; el intercambio de gases a través de la cáscara y de las membranas testáceas constituye la única vía respiratoria disponible para los huevos de las aves y de hecho para los de todos los vertebrados ovíparos [ver "Cómo respiran los huevos de las aves", de Herman Rahn, Amos Ar y Charles V. Paganelli; INVESTIGACIÓN Y CIENCIA, abril, 1979].

¿Qué podemos decir sobre el intercambio gaseoso cutáneo en el hombre? La piel humana no sirve, con seguridad, como órgano respiratorio de utilidad general para los tejidos corporales. Sin embargo, la piel, como cualquier tejido vivo, consume oxígeno y genera anhídrido carbónico. Todos los gases respiratorios consumidos o producidos por las células de la piel, así como una cantidad adicional no delimitada del anhídrido carbónico de la sangre capilar que circula a través de la piel, son intercambiados directamente entre ésta y el aire. De hecho, la permeabilidad de la piel es aprovechada en clínica para la administración de drogas por vía cutánea aplicando compresas.

A pesar de ser numerosos, los estudios sobre el intercambio gaseoso cutáneo han quitado importancia a la respiración a través de la piel en los vertebrados poniendo de manifiesto sus limitaciones. Por ejemplo, los investigadores han tendido a contrastar los pulmones y las branquias, elegantemente estructurados, con la piel. Han destacado las circunstancias que impiden el intercambio de gases a través de la piel y señalado, como características, las gruesas pieles, la presencia de escamas y el área limitada de la superficie cutánea, con el propósito de poner cortapisas al intercambio gaseoso cutáneo. Muchas de estas aseveraciones se basan en los principios físicos que rigen el intercambio gaseoso.

El oxígeno y el anhídrido carbónico atraviesan la membrana de un órgano respiratorio mediante un proceso físico denominado difusión. La difusión se define como el flujo de materia tendente hacia el equilibrio, mediante el movimiento al azar de las moléculas en un gas o en un líquido, desde una re-

gión de mayor concentración hasta otra de menor concentración. Debido a que las membranas de intercambio gaseoso de los pulmones y de las branquias son típicamente muy finas (menos de una milésima de milímetro en el pulmón humano), el equilibrio con el medio respiratorio se alcanza con suma rapidez. Una limitación más importante al intercambio gaseoso en los pulmones y en las branquias la impone la velocidad con que la sangre transporta el oxígeno desde la superficie de intercambio o aporta hacia ella el anhídrido carbónico. La rápida difusión posibilita que los animales con pulmones o branquias regulen el intercambio de gases simplemente aumentando o disminuyendo el flujo de sangre a través del órgano respiratorio.

En contraste, el incremento del flujo de sangre en la piel se cree que apenas incide sobre el intercambio gaseoso total en los vertebrados. Johannes Piiper, Peter Scheid, Randall Gatz y Eugene Crawford, del Instituto Max Planck de Medicina Experimental en Göttingen, han demostrado que el intercambio gaseoso cutáneo está limitado por la difusión. Es decir, que el proceso de difusión de los gases respiratorios a través de la piel, relativamente gruesa, de los vertebrados es tan lento, que no logra transferir el oxígeno y el anhídrido carbónico a la misma velocidad con que estos gases son transportados hacia o desde la piel por la sangre y por el medio respiratorio.

Otros estudios demostraron la existencia de otro problema relacionado con el intercambio gaseoso cutáneo. El equilibrio entre las concentraciones de los gases en la sangre que circula por los capilares de la piel y el medio respiratorio exterior está regido por la ley de Fick de la difusión: la tasa de intercambio gaseoso es proporcional a la diferencia entre la presión parcial del gas en el medio respiratorio y su presión parcial en la sangre. (La presión parcial, la presión de un gas en una solución o en una mezcla de gases, refleja la concentración del gas y su solubilidad.) Un gas se difundirá a través de la piel en la hipótesis exclusiva de que su presión parcial a un lado de la piel supere la del otro lado; cuanto mayor sea la diferencia se difundirá antes.

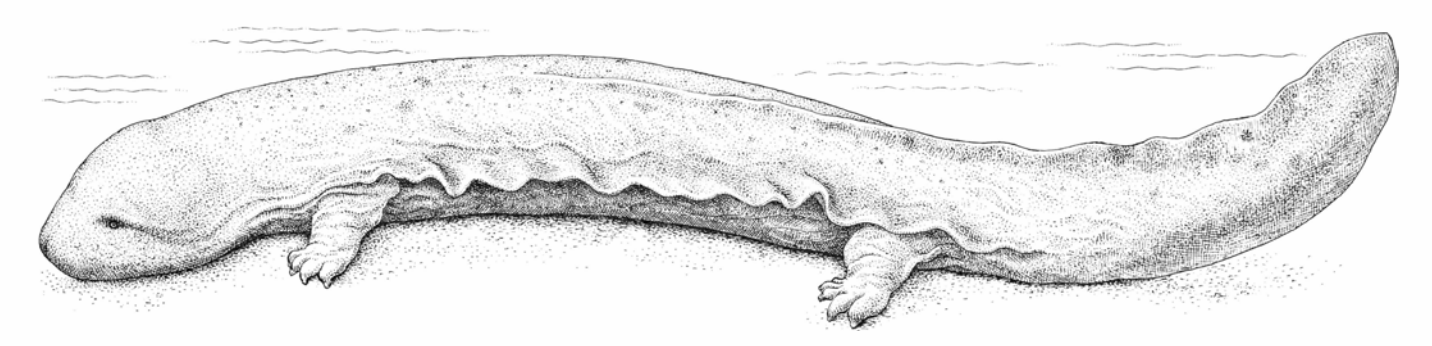
Como han demostrado los estudios de Donald C. Jackson y sus colegas de la Universidad de Brown, esta relación puede tener consecuencias importantes para los vertebrados que recurren a la respiración cutánea. Si el medio ambiente que rodea a un animal que respira a través de la piel tiene una presión parcial de anhídrido carbónico superior a la de la sangre del animal, éste ganará anhídrido carbónico procedente del medio ambiente, en vez de perderlo hacia él. De igual manera, si la elevación de la temperatura o la actividad determinan un incremento en la producción de anhídrido carbónico, muchos anfibios con respiración cutánea no pueden eliminar de forma inmediata el exceso de este gas, sino que deben esperar a que su presión parcial se eleve por encima de la exterior. La ley de Fick también rige la difusión del oxígeno. Nuestros propios estudios han probado que los anfibios pueden perder oxígeno cuando el nivel de este gas en su sangre supera al del agua en que se encuentran.

Los seres humanos pueden eliminar el exceso de anhídrido carbónico incrementando simultáneamente el flujo circulatorio pulmonar y la frecuencia respiratoria, mientras mantienen una presión parcial de anhídrido carbónico constante en la sangre. Sin embargo, las presiones parciales de gases internos de los animales con respiración cutánea parecen estar mal controladas y sujetas a la relación siempre variable entre las concentraciones externas de gas y las demandas respiratorias internas.

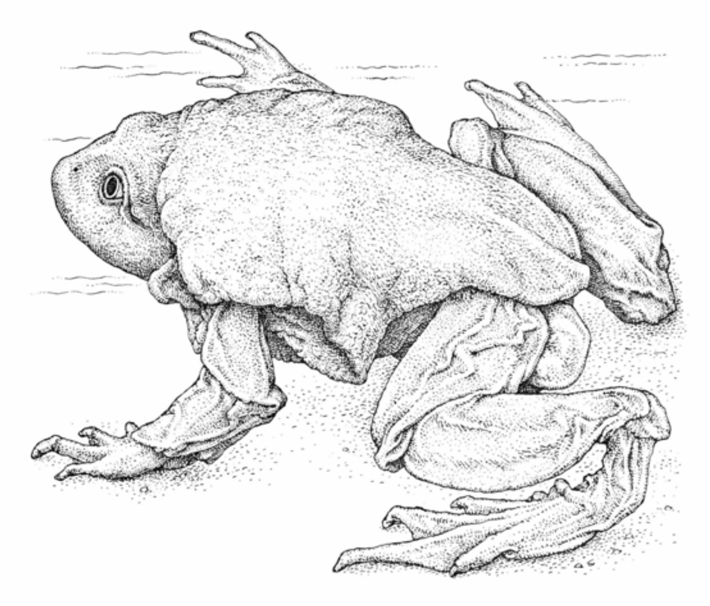
La cuestión es que, a pesar de las limitaciones manifiestas que el proceso de difusión impone, los vertebrados recurren al intercambio cutáneo de gas en una proporción significativa. Debemos reconocer que esta paradoja podría resolverse considerando los distintos mecanismos que los vertebrados tienen a su disposición para regular el intercambio gaseoso cutáneo. Algunos de estos mecanismos parecen evidentes, aunque rara vez se les ha reconocido su importancia al respecto. Otros de los procesos reguladores sugeridos son más sutiles. Los distintos mecanismos de regulación pueden dividirse en dos categorías: mecanismos que constituyen adaptaciones morfológicas permanentes o respuestas ante cambios a largo plazo (días, semanas o meses) en el medio ambiente o en la fisiología del animal y mecanismos a los que se recurre de un momento a otro a medida que las necesidades respiratorias inmediatas del organismo o las características del medio ambiente varían. En conjunto, estos mecanismos mantienen una capacidad efectiva considerable para controlar el intercambio gaseoso cutáneo.

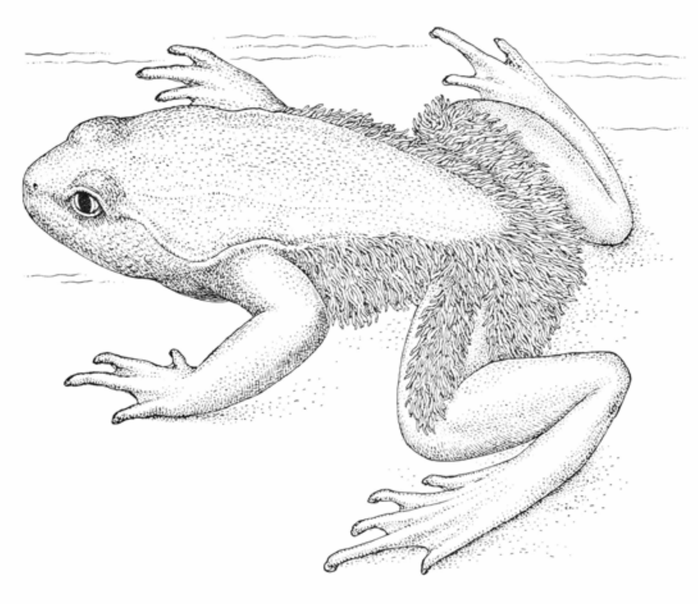
Un procedimiento evidente para regular la respiración a través de la piel está basado en el hecho de que el intercambio gaseoso cutáneo está en parte limitado por el área superficial total de la piel. En consecuencia, un cambio en este área puede aumentar o disminuir la cantidad de gas intercambiado a través de la piel. A pesar del esqueleto interno rígido y de la forma relativamente fija, muchos vertebrados presentan cambios estacionales del área superficial e incrementan el intercambio gaseoso cutáneo precisamente mediante estos cambios.

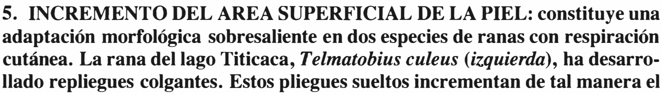
Por ejemplo, algunos anfibios macho se entregan a rituales de apareamiento elaborados e intensos que incluyen, a veces, movimientos corporales repetidos durante horas. En relación con este



4. DIABLO DEL LODO (*Cryptobranchus allegiensis*). Es una gran salamandra acuática que se encuentra en los cursos rápidos del este y del centro de los Estados Unidos. Alcanza una longitud de unos 70 centímetros. Aunque posee pulmones, respira principalmente a través de su piel, que posee numerosos pliegues.







5. INCREMENTO DEL AREA SUPERFICIAL DE LA PIEL: constituye una adaptación morfológica sobresaliente en dos especies de ranas con respiración cutánea. La rana del lago Titicaca, *Telmatobius culeus* (*izquierda*), ha desarrollado repliegues colgantes. Estos pliegues sueltos incrementan de tal manera el área dérmica para el intercambio gaseoso que la rana no necesita respirar en absoluto a través de sus pulmones. En el sapo peludo, *Astylosternus robustus* (*derecha*), los machos presentan papilas dérmicas en la parte posterior del cuerpo, que utilizan como órganos accesorios en el intercambio gaseoso.

comportamiento se produce un gran aumento en la demanda de oxígeno y en la eliminación de anhídrido carbónico. En respuesta, aparentemente, a la sobrecarga respiratoria, partes de la piel de estos anfibios se amplían o desarrollan evaginaciones durante la época de celo. Estas superficies actúan de órganos respiratorios accesorios.

Cambios en la estructura de la piel con esta finalidad aparecen en el macho de la rana peluda, *Astylosternus robustus*. Finas papilas dérmicas que superficialmente parecen pelo se ponen de manifiesto en la parte posterior del cuerpo. La aleta caudal crecida y la cresta dorsal que se desarrolla en los machos de muchas especies de tritones durante la época de reproducción podrían también contribuir a incrementar la cantidad total de gas que difunde a través de la piel. El aumento de superficie de la piel de los machos experimenta, de forma característica, una regresión al final de la época de reproducción. Además, en las hembras, relativamente pasivas, de estas especies no se dan en absoluto este tipo de estructuras.

Amén de estos cambios estacionales en el área superficial de la piel, los anfibios han desarrollado una serie de características morfológicas permanentes que promueven la respiración cutánea. Muchos anfibios tienen cuerpos o colas desproporcionadamente alargados en apariencia. Se ha argumentado que estas formas les dotarían de piel suficiente para que se produjera un intercambio gaseoso cutáneo adecuado.

Otras especies de anfibios están provistas permanentemente de numerosos pliegues de la piel que también incrementan el área superficial para el intercambio gaseoso cutáneo. La más espectacular de estas especies es la rana del lago Titicaca, *Telmatobius culeus*, que habita en ese lugar andino. Como descubrieron Victor Hutchison y sus colaboradores en la Universidad de Oklahoma, la rana del lago Titicaca se halla perfectamente adaptada a la respiración cutánea y no necesita ventilar sus pulmones en absoluto. La causa fundamental de esta capacidad la constituyen los pliegues colgantes de la piel que sobresalen en el tronco y patas posteriores. Otras especies del género *Telmatobius* y también de otros géneros de ranas poseen pliegues cutáneos similares. En algunos tipos de anfibios, como en el diablo del lodo, *Cryptobranchus alleganiensis*, una gran salamandra acuática, los pliegues cutáneos son pequeños, numerosos y con una rica vascularización capilar.

Una segunda adaptación que aumenta la capacidad para el intercambio gaseoso cutáneo consiste en la reducción del espesor de la piel para reducir su resistencia a la difusión de los gases respiratorios. La piel de los anfibios, además de estar desprovista de barreras físicas como las que representan el pelo, las plumas o las escamas, posee, de modo característico, un grosor de sólo entre 10 y 50 micrometros (millonésimas de metro). En realidad, el factor morfológico significativo que rige la difusión en este aspecto no es el espesor total de la piel, sino el de la "barrera de difusión", la distancia entre el medio respiratorio externo a la piel y la sangre que circula por los capilares cutáneos. Por consiguiente, cualquier cambio morfológico que reduzca la distancia entre el torrente circulatorio y el medio respiratorio facilita un mayor intercambio de gases.

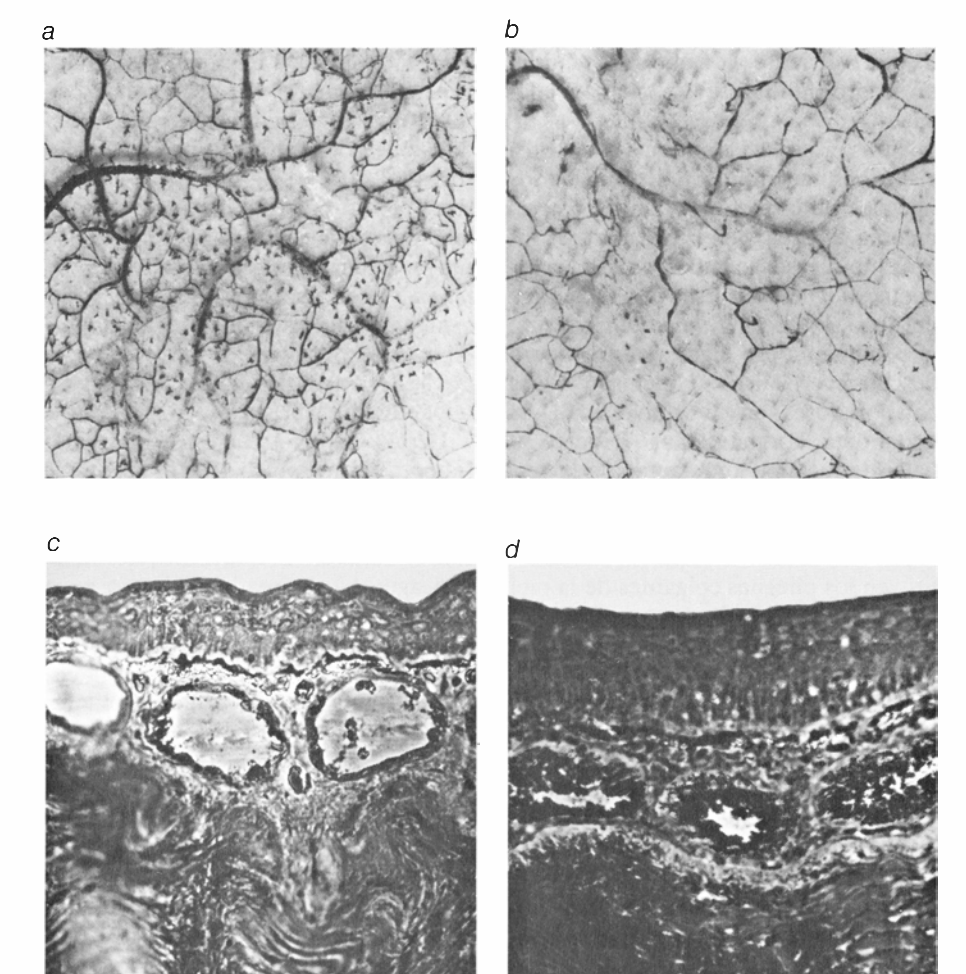
Demostramos a través de un experimento sencillo que estos cambios pueden ocurrir en una fracción de la vida de un anfibio. Se mantuvieron larvas de rana en dos recintos; contenía uno agua bien aireada y el otro agua con una concentración de oxígeno de aproximadamente la mitad del anterior. Cuatro semanas más tarde, la barrera de difusión de las larvas mantenidas en el agua bien aireada era de 40 micrometros, un valor típico. En contraste, la barrera de difusión alcanzó un promedio de sólo 20 micrometros en los animales albergados en el agua pobre en oxígeno. Además, la red capilar cutánea de las larvas privadas del oxígeno normal resultó más densa y refinada. Por lo que se vio, estas larvas experimentaron cambios adaptativos de aclimatación que aumentaron su capacidad para el intercambio gaseoso cutáneo.

Incluso los vertebrados con piel relativamente gruesa pueden seguir desarrollando un intercambio gaseoso sig-

nificativo, siempre que la evolución haya conducido a un emplazamiento favorable de los capilares cutáneos. Por ejemplo, en la piel escamosa de algunos lagartos, los capilares cutáneos se localizan bien sea bajo el epitelio de entre las escamas o bien discurren bajo las articulaciones de las escamas, donde éstas son más finas. En algunas serpientes los capilares cutáneos penetran en la propia escama. Las escamas dérmicas de los peces están recubiertas generalmente por una capa de tejido vivo, una disposición que coloca a los capilares cutáneos sobre la escama resistente a la difusión y muy próximos al medio respiratorio.

Las adaptaciones como son los pliegues cutáneos, papilas dérmicas y, de forma general, la piel fina, así como la circulación especializada que abastece a estas estructuras, se desarrollan obviamente con lentitud, si no se comparan con una escala evolutiva de tiempo, durante días o semanas. ¿Cómo un vertebrado puede regular el intercambio gaseoso cutáneo de forma instantánea si su demanda de oxígeno aumenta de súbito (como ocurre al inicio de la actividad) o si de improviso se encuentra con una zona de agua de concentración elevada de anhídrido carbónico?

En la mayoría de vertebrados que hemos examinado, no todos los capilares cutáneos son perfundidos con sangre de manera continua. La piel que se encuentra alejada de los capilares subyacentes o que tiene en su base capilares no perfundidos no contribuye al intercambio gaseoso total; el área superficial funcional de la piel en un momento dado consiste sólo en las regiones de la piel que recubren capilares cutáneos perfundidos.



6. ADAPTACIONES MORFOLOGICA MICROSCOPICAS. En las cuatro fotografías se muestran las respuestas microscópicas que favorecen la respiración cutánea bajo condiciones adversas. La red capilar de la piel de las larvas de rana es más refinada y densa cuando éstas viven en agua con bajo contenido de oxígeno (*a*) que cuando se desarrollan en agua rica en este gas (*b*). La distancia entre los capilares y la superficie de la piel en las larvas puede también variar de forma dependiente del contenido en oxígeno del agua: los capilares de las larvas desarrolladas en agua poco oxigenada se encuentran más próximos a la superficie de la piel (*c*) que los capilares de las larvas mantenidas en agua donde hay alta concentración de oxígeno (*d*).

El inicio del flujo de sangre a través de los capilares (llamado reclutamiento capilar) puede ocurrir en el orden de segundos. (El rubor es un ejemplo clásico, aunque no relacionado con la respiración, de este fenómeno en el hombre.) El reclutamiento capilar en la piel de los anfibios fue ilustrado hace muchos años por Piotr Pocozpko y sus colaboradores en Polonia. Encontraron que en ranas que respiraban una concentración elevada de anhídrido carbónico y en ranas a las que se impedía utilizar sus pulmones, el número de capilares cutáneos perfundidos aumentaba en una proporción aproximada de un tercio.

Una confirmación experimental adicional de que el reclutamiento capilar guarda relación con el intercambio gaseoso cutáneo ha surgido a partir de nuestros estudios en ranas toro. Mientras trabajábamos en la Universidad de Massachusetts en Amherst, observamos que, cuando las ranas sumergidas en el agua quedaban expuestas al aire, el número de capilares cutáneos perfundidos descendía rápidamente en un sesenta por ciento; al propio tiempo la eliminación de anhídrido carbónico a través de la piel bajaba en un 44 por ciento. Cuando las ranas eran devueltas al agua, se producía reclutamiento capilar y la eliminación cutánea de anhídrido carbónico retornaba inmediatamente a su nivel normal.

Experimentos más recientes llevados a cabo por Gary Malvin y Michael P. Hlastala, de la Facultad de Medicina de la Universidad de Washington, han demostrado también que las ranas controlan el flujo de sangre capilar en su piel. Estas investigaciones pusieron de manifiesto que las ranas reducían la pérdida de gas a través de la piel en una proporción del 15 al 30 por ciento, cuando se las exponía a una atmósfera carente de oxígeno, probablemente a través de una disminución en la proporción de piel perfundida por la sangre.

Tan importantes como las adaptaciones morfológicas y el flujo sanguíneo capilar son los procesos fisiológicos o las respuestas de comportamiento que afectan a las presiones parciales de los gases respiratorios en la sangre del interior de la piel o en el medio respiratorio externo a ésta. Los más importantes son los procesos que regulan el flujo tanto del medio respiratorio como de la sangre a lo largo de la barrera de difusión.

La ventilación, el flujo del medio

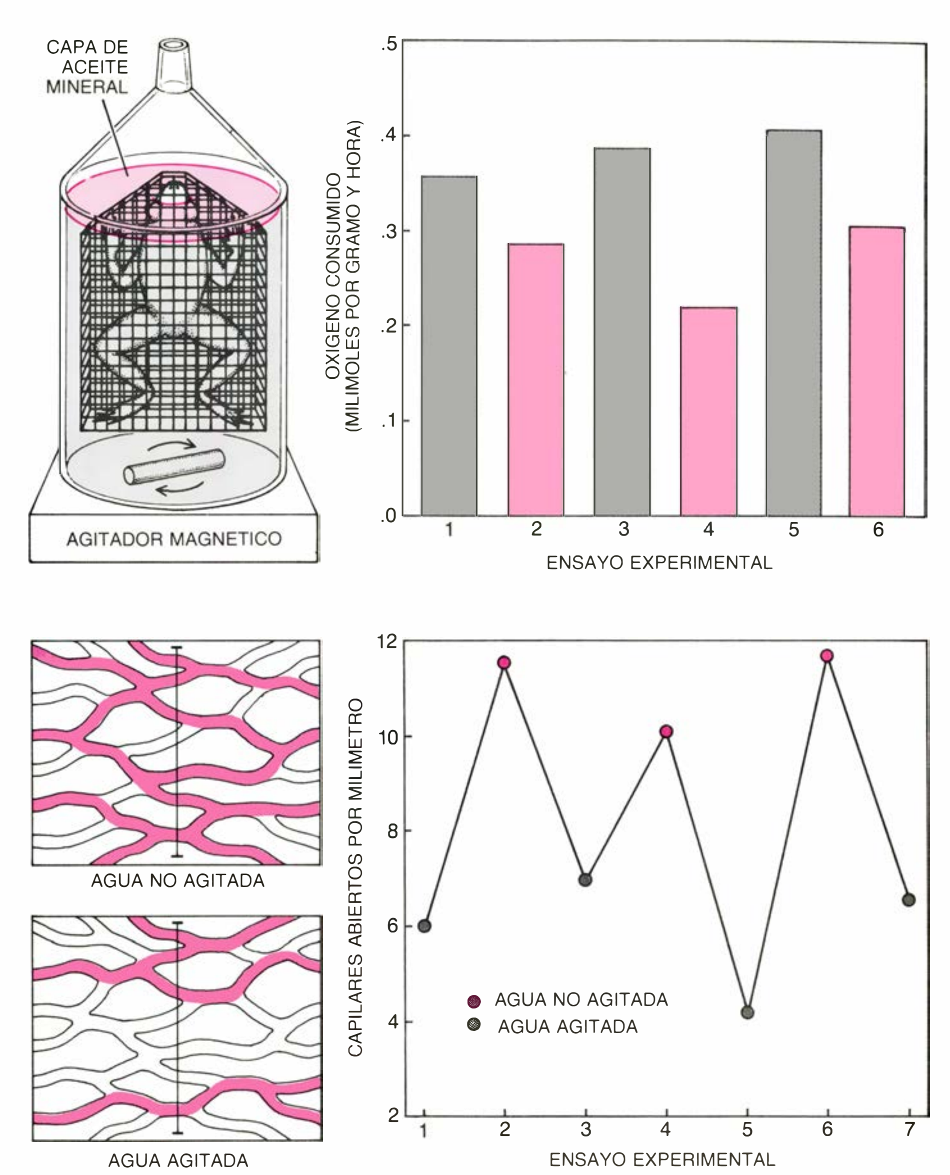
respiratorio hacia el órgano respiratorio o a través de éste, es un factor manifiestamente crítico en los pulmones y en las branquias internas. Si la ventilación se detiene, el intercambio gaseoso decae bruscamente, porque, en el medio respiratorio retenido en el cuerpo, el oxígeno se agota en seguida (su presión parcial disminuye) y el anhídrido carbónico se acumula muy pronto (su presión parcial aumenta). Por otra parte, la ventilación se ha considerado a menudo innecesaria en el contexto del intercambio gaseoso cutáneo; después de todo, la piel de un vertebrado situado en el aire o en una gran extensión de agua está en contacto permanente con un "depósito infinito" de medio respiratorio que contiene abundantes reservas de oxígeno y únicamente pequeñas cantidades de anhídrido carbónico.

Sin embargo, la rana del lago Titicaca ondea sus grandes pliegues cutáneos y el diablo del lodo balancea su cuerpo. Además, ambos animales aumentan la frecuencia de estos movimientos cuando disminuye la concentración de oxígeno en el agua. ¿Puede resultar útil este comportamiento para aumentar el intercambio gaseoso cutáneo o, por el contrario, no guarda ninguna relación con la respiración?

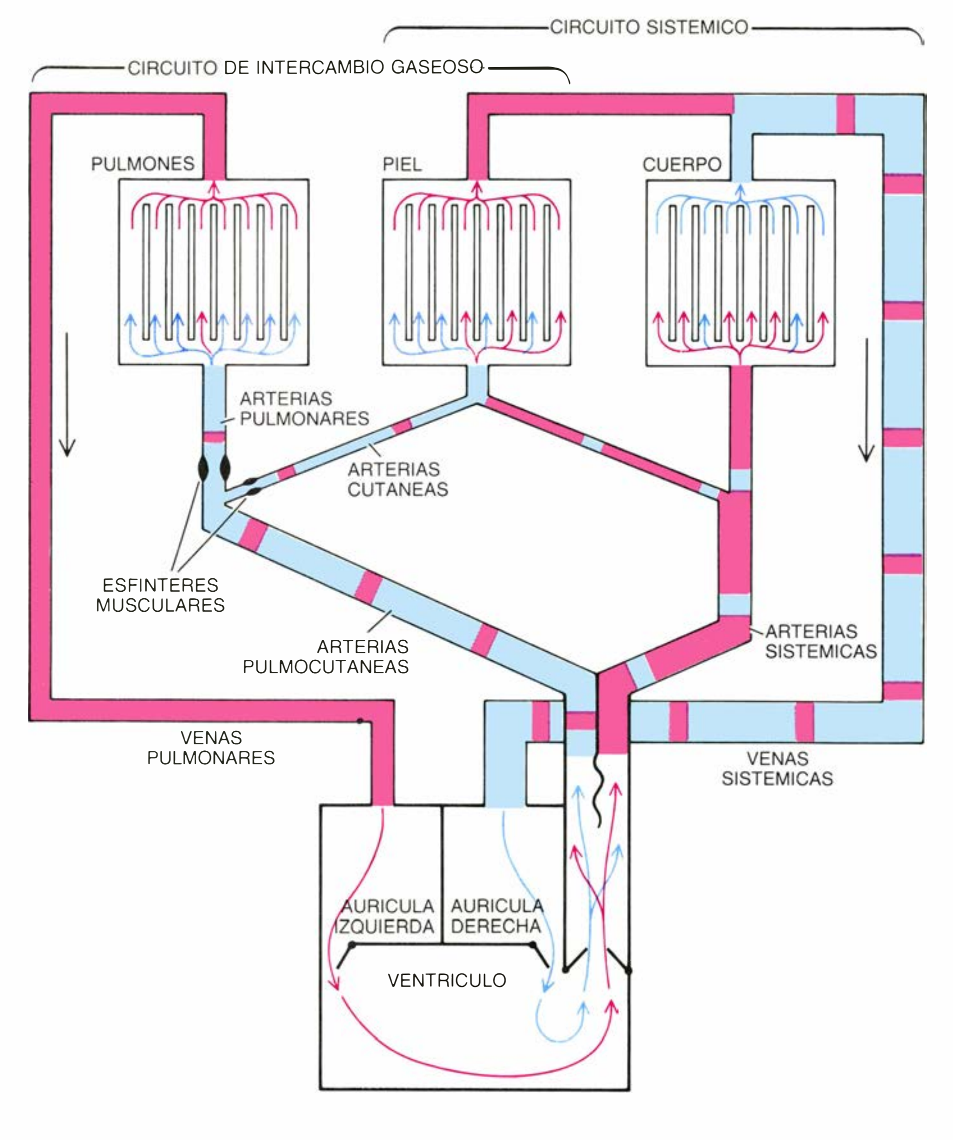
El examen de este peculiar comportamiento nos llevó a reflexionar sobre los aspectos físicos que intervienen en el intercambio de calor. El intercambio térmico está a menudo muy restringido tanto en el aire como en el agua; se ve favorecido cuando el medio (o la fuente de calor) se mueve. Por ejemplo, si nos metemos en una bañera con agua caliente y permanecemos inmóviles, sentiremos dolor a medida que el calor penetra en la piel. Al enfriarse la capa de agua inmediata a la piel, el fluido forma una capa límite algo más fría. Cualquier movimiento rápido de la piel o del agua disipará la capa límite, permitiendo, otra vez, al agua caliente entrar en contacto con la piel y causando dolor de nuevo, hasta que la capa se reestablezca. Incluso en un depósito infinito (o en la bañera limitada) de agua caliente, el estancamiento del medio contiguo a la piel puede restringir la transferencia de calor.

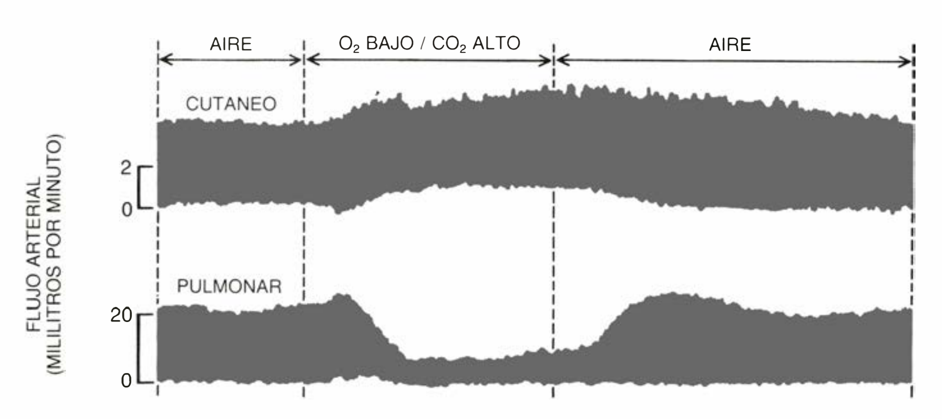
Consideramos que se produce una situación análoga para el intercambio de gases respiratorios, particularmente en el agua. Si tanto el animal como el agua están en situación estacionaria, el oxígeno que se difunde desde el agua adyacente a la piel hasta el torrente circulatorio crearía una capa límite de difusión de baja presión parcial de oxígeno. Debido a que la tasa de difusión es proporcional a la diferencia entre las presiones parciales de oxígeno a un lado y otro de la barrera de difusión, el intercambio gaseoso cutáneo disminuiría. El movimiento, bien sea del animal, del agua o de ambos, podría disipar la capa límite y de este modo incrementar la difusión de oxígeno a través de la piel. Calculamos que la capa límite de difusión ofrece una resistencia significativa al intercambio gaseoso a velocidades de flujo de agua de cuatro centímetros por segundo o menores. Bastarían, sin embargo, velocidades muy inferiores para disipar la capa límite de difusión en el aire.

Para someter a prueba nuestra hipótesis, inmovilizamos ranas toro entre dos rejillas metálicas; ello nos aseguraría que los movimientos corporales espontáneos no disiparían la capa límite y que su área superficial cutánea sería constante y máxima. Colocamos entonces a cada animal, intercalado en-



7. VENTILACION DE LA PIEL: afecta al intercambio gaseoso cutáneo. Se sumergió una rana, inmovilizada entre dos rejillas metálicas, en una cámara llena de agua, de forma que tan sólo asomaran a la superficie los orificios nasales (*arriba, izquierda*). Poniendo en marcha un agitador situado en el fondo de la cámara se pudo ventilar la piel. Se determinó la captación cutánea de oxígeno a partir del descenso en la concentración de este gas en la cámara. (Una capa superficial de aceite mineral evitaba la aireación del agua.) El intercambio de oxígeno a través de la piel (*arriba, a la derecha*) fue superior al agitar el agua (*barras grises*) que al permanecer ésta en calma (*barras de color*). Se demostró que la reducción observada en el consumo de oxígeno con el agua en calma no dependía de una disminución en el flujo de sangre capilar. Se colocó la pata de una rana sobresaliendo por el lado del recipiente de manera que se pudieran observar los capilares de las membranas interdigitales a través del microscopio. Se contó el número de capilares con sangre circulante situados en intersección con una determinada línea del ocular del microscopio mientras el agua permanecía en calma y al agitarla (*abajo, izquierda*). El número de capilares perfundidos por la sangre aumentaba si no se agitaba el agua; dicho de otra manera: en ausencia de ventilación cutánea (*parte inferior, a la derecha*).





8. SISTEMA CIRCULATORIO DE LOS ANFIBIOS ilustrado en esquema. Permite que la sangre desoxigenada (*azul*) alcance la piel y por consiguiente favorece la difusión de oxígeno a través de la piel hacia la sangre. La mayor parte de la sangre desoxigenada que llega a la piel proviene del corazón a través de las arterias cutáneas (*parte superior*); una pequeña cantidad se mezcla con sangre oxigenada (*rojo*) en el ventrículo del corazón y es transportada por las arterias sistémicas hacia la piel. (También penetra algo de sangre oxigenada en las arterias pulmocutáneas.) La separación de la sangre se mantiene a pesar de la existencia de un solo ventrículo que impulsa tanto a la sangre oxigenada como a la desoxigenada hacia las arterias: la mayor parte de sangre oxigenada es dirigida principalmente hacia el circuito sistémico y la desoxigenada principalmente hacia el circuito de intercambio gaseoso. La sangre predominantemente desoxigenada de las arterias pulmocutáneas puede desviarse mayoritariamente bien hacia los pulmones o bien hacia la piel, dependiendo de qué órgano se encuentre en un momento determinado en mejores condiciones para la respiración. El incremento que se produce en el flujo cutáneo y el descenso en el flujo pulmonar cuando los pulmones se llenan con un gas rico en anhídrido carbónico y pobre en oxígeno puede ser registrado mediante reómetros electromagnéticos en las arterias cutáneas y en las arterias pulmonares (*esquema inferior*). Los esfínteres musculares situados en las arterias pueden regular el flujo actuando a modo de válvulas.

tre las rejillas, en una cámara estanca. Se llenó la cámara con agua de tal manera que únicamente los orificios nasales de la rana asomaran a la superficie. Pudimos medir el descenso en la concentración de oxígeno en los compartimentos de aire y de agua y, en consecuencia, calcular los respectivos consumos de oxígeno pulmonar y cutáneo. Poniendo en marcha un agitador en el fondo de la cámara se pudo ventilar la piel con el agua del recipiente.

Los resultados obtenidos sustentan nuestra hipótesis. Cuando detuvimos la ventilación de la piel, el consumo de oxígeno cutáneo (determinado a partir de la concentración de oxígeno en el agua) disminuyó en una proporción aproximada de un tercio. Este resultado contradice manifiestamente la idea según la cual la ventilación carece de importancia en el intercambio gaseoso cutáneo.

Como hemos destacado, la magnitud del intercambio gaseoso cutáneo viene determinada por diferentes variables, además de la ventilación; verbigracia, el área superficial funcional de la piel y la presión parcial de oxígeno en la parte interna de la barrera de difusión. ¿Podría, de hecho, ser responsable alguno de estos otros factores o la capa límite de la disminución del consumo de oxígeno cutáneo asociado a la detención de la ventilación?

Nuestros experimentos anteriores nos habían señalado la importancia del área superficial funcional y del reclutamiento capilar. En consecuencia, repetimos los ensayos de ventilación en ranas observando las proporciones relativas de capilares perfundidos y no perfundidos. La cantidad de oxígeno captado a través de la piel podría haber disminuido mediante una reducción en el número de capilares perfundidos en la piel de las ranas. Nos encontramos con que ocurría lo contrario: cada vez que deteníamos la ventilación, se producía un reclutamiento adicional de capilares cutáneos. Dado que el reclutamiento capilar debería incrementar el intercambio gaseoso cutáneo, la disminución en la captación de oxígeno a través de la piel, que se observa siempre que se suspende la ventilación, no puede deberse a cambios en el área funcional superficial.

En un tercer experimento medimos la presión parcial del oxígeno en la sangre de las arterias que se dirigen hacia la piel. Si la presión parcial del oxígeno aumentara cada vez que la ventilación cesa, este hecho podría explicar el des-

censo observado en la proporción del oxígeno que se difunde desde el agua hasta la sangre. Sin embargo, la agitación del agua no modificó la presión parcial del oxígeno de la sangre que circulaba por los citados vasos. Los resultados sugieren que la formación de una capa límite de difusión era responsable del descenso en la captación cutánea de oxígeno y que la ventilación de la piel servía para regular el intercambio gaseoso con la disipación de esta capa.

Las presiones parciales del oxígeno y del anhídrido carbónico en la sangre de los capilares cutáneos varía dependiendo de si se trata o no de sangre oxigenada. Esto constituye otro posible mecanismo de regulación. Controlando si la sangre que circula hacia la piel es mayoritariamente oxigenada o desoxigenada, un organismo podría regular, a buen seguro, la cantidad de oxígeno y de anhídrido carbónico que se difunden a través de la piel.

El intercambio gaseoso cutáneo alcanzaría un grado de mayor efectividad si sólo circulara hacia la piel sangre desoxigenada, de la misma manera que únicamente este tipo de sangre circula hacia los pulmones en los mamíferos. Las diferencias entre las presiones parciales a un lado y otro de la piel, tanto del oxígeno como del anhídrido carbónico, serían las mayores en estas circunstancias. Sin embargo, la piel de la mayoría de los vertebrados no se diferencia, a este propósito, de cualquier otro tejido: sólo recibe sangre procedente de las principales arterias sistémicas, que suministran de forma característica sangre oxigenada. Como la diferencia entre la presión parcial del oxígeno de la sangre oxigenada y la del medio respiratorio suele ser muy pequeña, la captación de oxígeno del medio ambiente está normalmente limitada por el propio tipo de sangre que las células epiteliales necesitan para vivir. Incluso bajo ésta, en apariencia, principal restricción, la piel puede, sin embargo, tener importancia en la eliminación de anhídrido carbónico porque los niveles de este gas en la sangre arterial pueden todavía seguir siendo considerablemente superiores a los del medio ambiente. Esto explica en parte por qué motivo la eliminación de anhídrido carbónico sobrepasa, de forma característica, la captación de oxígeno en el intercambio gaseoso cutáneo entre los vertebrados.

De hecho, algunos vertebrados dirigen parte de su sangre desoxigenada hacia las arterias sistémicas para así aumentar el intercambio gaseoso cutáneo. Los anfibios y los reptiles, por ejemplo, poseen un corazón con tabicación incompleta, lo que permite que la sangre desoxigenada se dirija hacia la piel sin atravesar antes los pulmones. Tradicionalmente, los anatomistas comparados han considerado esta disposición como primitiva e ineficiente. Kjell Johansen, de la Universidad de Aarhus, y Fred White, de la Institución de Oceanografía Scripps, entre otros, han sugerido que lo cierto es todo lo contrario. Han señalado que la estructura del corazón de los anfibios y los reptiles constituye una importante adaptación que les permite distribuir la sangre de la manera que mejor promueva el intercambio gaseoso.

Además, los anfibios presentan la singularidad de tener arterias cutáneas que permiten transferir directamente sangre desoxigenada hacia la piel. La sangre puede abandonar el ventrículo único del corazón de un anfibio por cualquiera de dos vías. Una a través de las arterias sistémicas, que transportan sangre oxigenada directamente hacia el cerebro, los músculos, las vísceras y finalmente hacia la piel. La segunda vía es a través de las arterias pulmocutáneas, que suministran sangre desoxigenada a los pulmones por medio de las arterias pulmonares y a la piel por medio de las arterias cutáneas.

Graham Shelton y sus colegas, de la Universidad de East Anglia, han demostrado que los anfibios pueden dirigir la sangre desoxigenada hacia las arterias pulmocutáneas y desde allí selectivamente bien hacia el pulmón o hacia la piel para el intercambio gaseoso. La base de esta capacidad puede radicar en la estructura del corazón y en los esfínteres musculares que rodean a las arterias pulmonares y cutáneas, después de separarse éstas a partir de la arteria pulmocutánea común. Los esfínteres pueden actuar de válvulas que desvían el flujo de sangre por una u otra vía. En cualquier caso, la distribución exacta del flujo depende de los niveles prevalecientes de oxígeno y de anhídrido carbónico en estos dos órganos respiratorios.

Trabajando con Nigel West, de la Universidad de Saskatchewan, determinamos la distribución de la sangre pulmocutánea entre la piel y los pulmones implantando transductores electromagnéticos de flujo alrededor de las arterias pulmonares y cutáneas de sapos anestesiados. Cuando simulamos las presiones parciales de oxígeno y de anhídrido carbónico que se dan característicamente en los pulmones de los sapos en condiciones de apnea (aguantando la respiración), los sapos desviaron la sangre desde los pulmones hacia la piel. Esta respuesta facilita el intercambio gaseoso cutáneo.

Este tipo de reacción debe producirse también en las ranas siempre que la respiración pulmonar se vea obstaculizada, como sucede cuando la rana se encuentra debajo del agua. Recientemente, Robert Boutelier, Mogens Glass y Norbert Heisler, del Instituto Max Planck en Göttingen, realizaron una serie de experimentos, similares a los nuestros, en ranas toro no anestesiadas. Inyectaron esferas microscópicas radiactivas en los sistemas circulatorios de las ranas toro. Las esferas, de un tamaño ligeramente superior al de los glóbulos rojos, se alojaban en los capilares a través de los que fluía la sangre. Se pudo calcular la distribución del flujo de sangre entre los pulmones, la piel y los otros tejidos corporales a partir de la emisión radiactiva de los distintos tejidos expuestos mediante disección.

Cuando las ranas se sumergían en agua rica en oxígeno después de respirar mezclas de gases con baja concentración de oxígeno, la sangre pulmocutánea se distribuía con preferencia hacia la piel en lugar de hacia los pulmones. Recíprocamente, cuando las ranas se sumergían con sus pulmones llenos de aire en agua pobre en oxígeno, la distribución de la sangre pulmocutánea favorecía a los pulmones y no a la piel. Los anfibios pueden, de forma manifiesta, regular el flujo cutáneo de sangre tanto para optimizar el intercambio gaseoso cutáneo como para coordinar la actividad respiratoria de la piel con la de los pulmones.

Aunque la respiración cutánea puede ser responsable de sólo una pequeña fracción del intercambio gaseoso total en algunos animales, en otros puede desempeñar un papel principal o incluso vital. La amplia diversidad de especies que recurren, al menos en parte, al intercambio gaseoso cutáneo debería bastar para convencernos de que la respiración cutánea no constituye un hecho excepcional, sino habitual en los vertebrados. A partir de una investigación más atenta, el intercambio gaseoso cutáneo se ha manifestado como un proceso bien regulado, energéticamente económico y capaz de responder ante los cambios inmediatos, prolongados y evolutivos en las demandas respiratorias de un animal.

Vuelo de propulsión humana

Las aeronaves de propulsión humana utilizan un régimen de vuelo poco conocido. Las que pueden afrontarlo son de vuelo agradable y se emplearán en servicios de reconocimiento y en ciencias planetarias

Mark Drela y John S. Langford

Desde que la sociedad empezó a soñar con el vuelo humano, suponíase, siempre, que el volador aportaría la potencia. Así lo hacen los pájaros, a quienes se pretendía emular. Sólo en los últimos 25 años –tras el desarrollo del avión de hélice, el reactor, el vuelo supersónico y el espacial– han hecho valer sus méritos las aeronaves de propulsión humana. Su aparición débese al desarrollo de un conjunto de técnicas cruciales: aerodinámicas, propulsivas y estructurales. Igualmente importante fue un logro algo más reciente: la fabricación de naves gobernables por un piloto sobre quien recae además la tarea de generar una gran cantidad de potencia mecánica. Ha llegado ya el momento en pensar cómo sacarles partido en ciertas aplicaciones.

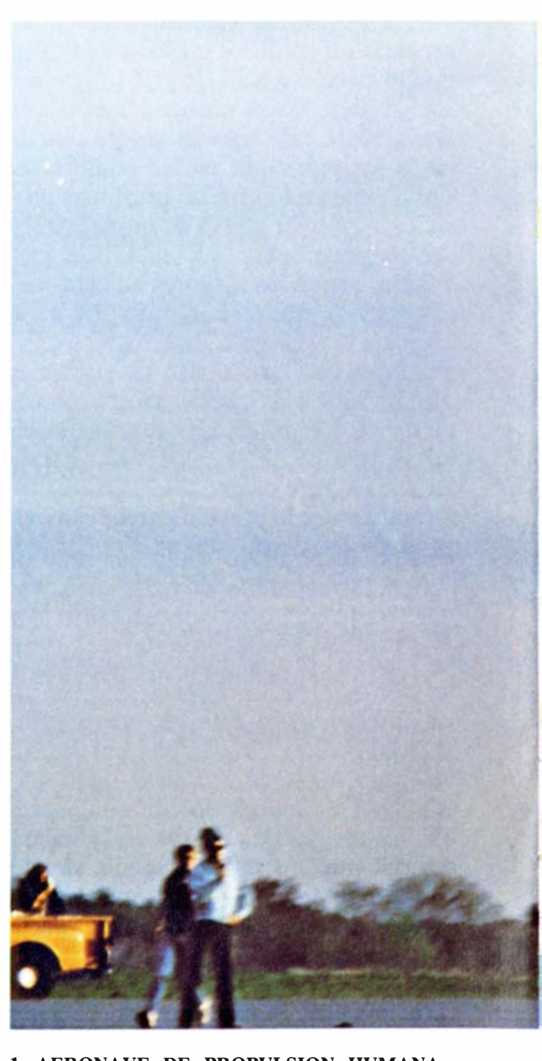
Probablemente, las aeronaves de propulsión humana no habrían alcanzado esta etapa sin el estímulo proporcionado por una serie de competiciones patrocinadas por ciertas organizaciones y algunas personas. La primera, que tuvo lugar en Francia entre 1912 y 1922, era un proyecto de la compañía Peugeot, a cuyas instancias aparecieron aeronaves que en realidad sólo eran bicicletas saltadoras: el piloto pedaleaba fuerte para conseguir velocidad en el suelo; estos aparatos con alas lograban planear en el aire una docena de metros. Pero instalada ya la nave en el aire, carecía de medios de propulsión.

En 1935, la aeronave alemana *Mufli* dio un paso adelante: el piloto llegaba a mover una hélice tras el despegue desde una catapulta. Evidentemente, las necesidades de potencia para mantener el vuelo eran mayores que las alcanzables entonces por los diseñadores. El piloto apenas podía producir potencia suficiente para un largo planeo, el mayor de los cuales fue de 712 metros. El *Mufli* se presentó a una competición en la que un grupo de Frankfurt ofrecía 5000 marcos de premio al primer vuelo de propulsión humana que uniera dos pilones distantes 500 metros. Premios similares se ofrecieron también en Italia y la Unión Soviética. Todos quedaron desiertos, sin embargo.

Las competiciones más famosas, y las que auspiciaron un auténtico progreso tecnológico, fueron las patrocinadas por el industrial británico Henry Kremer. En 1959 ofreció un premio de 5000 libras al primer participante que pudiera hacer volar una aeronave durante una milla en un trayecto en forma de ocho, sólo con propulsión humana. Habían pasado dieciocho años y el premio en metálico se había multiplicado por diez cuando Bryan Allen, de Estados Unidos, voló con éxito en el *Gossamer Condor* a lo largo de dicho trayecto.

Kremer ofreció luego el mayor premio de la historia de la aviación: 100.000 libras para el primer vuelo de propulsión humana a través del Canal de la Mancha. De nuevo el ganador fue Allen, que pedaleó el *Gossamer Albatross* a través de las 21 millas entre Folkestone y el cabo Griz-Nez, el 12 de junio de 1979.

Tanto el *Condor* como el *Albatross* eran naves grandes y frágiles que se volvían ingobernables si lo que soplaba no era una ligera brisa. Sus éxitos no despertaron mayor interés en el vuelo de propulsión humana. Cuatro años después del cruce del Canal, Kremer tornó a patrocinar otra competición tendente a que las aeronaves de propulsión humana fueran más rápidas y, por tanto, prácticas y menores. Esta vez se trataba de alcanzar una velocidad relativamente alta en un trayecto triangular de 1500 metros. Se concedía un premio de 20.000 libras para el primero que terminara el trayecto en menos de tres minutos –una marcha que suponía una velocidad de unos 32 kilómetros por hora. El estadounidense Frank P. Scarabino se alzó con la victoria en mayo de 1984, pilotando el *Monarch*, una nave diseñada y construida en el Instituto de Tecnología de Massachusetts (MIT). Actualmente



1. AERONAVE DE PROPULSION HUMANA realizando una prueba de velocidad en Hanscom Field, en el estado de Massachusetts. Se trata de *Monarch B*, nave diseñada y construida en el Instituto de Tecnología de Massachusetts. La pilota Frank P.

hay premios ofrecidos por la Royal Aeronautical Society para vuelos que mejoren el récord existente en al menos un 5 por ciento; tres de ellos ya se han adjudicado.

Sin contar los aparatos de acrobacia y las bicicletas aladas del cambio de siglo, se han construido unas 60 aeronaves de propulsión humana, la mayoría de ellas pensando en las competiciones de Kremer. Los diseños pueden agruparse a grandes rasgos en tres generaciones, según las características aerodinámicas y los diseños estructurales [*véase la figura 3*]. Las naves de la primera generación dependían, en sus supuestos teóricos, de los planeadores sin motor. Sólo podían volar en línea recta, y pocas veces superaban el kilómetro.

La segunda generación comprende los primeros vehículos capaces de vuelo sostenido y controlable de propulsión humana. El *Gossamer Condor* es el más conocido de ellos. Fue construido en California por un equipo que dirigía Paul MacCready, Jr.; se halla expuesto en el Museo Nacional del Aire y el Espacio de la Institución Smithsoniana. Una nave menos conocida, pero algo más robusta, es la *Chrysalis*, un biplano construido por estudiantes del MIT en 1979. El aspecto que presentan los aparatos de la segunda generación es sorprendente, puesto que los diseñadores dejaron atrás las ideas convencionales acerca de la forma que debía tener un aeroplano.

Las aeronaves de tercera generación se han construido para competiciones de velocidad. Son, pues, mucho menores. Aunque en su forma externa recuerdan todavía las de la primera generación, incorporan moderna tecnología aerodinámica y estructural y reflejan la experiencia acumulada por el diseño y funcionamiento de las máquinas de segunda generación. Debido a una disposición de las normas de las competiciones de velocidad, algunas de estas máquinas tienen también capacidad para almacenar energía, mediante sistemas que permiten al piloto hacer acopio de su propia energía en la nave durante un corto tiempo antes del vuelo, Lo que generalmente se consigue pedaleando para cargar las baterías mediante un generador. Estas máquinas han adquirido una complejidad y refinamiento muy superiores a los de sus predecesoras. Por ejemplo, el *Monarch*, del MIT, incorpora un dispositivo electrónico con el que se regula la velocidad de la hélice, lo que modifica las necesidades del ritmo de pedaleo o la salida de potencia almacenada.

Aunque las condiciones específicas de cada competición de Kremer han

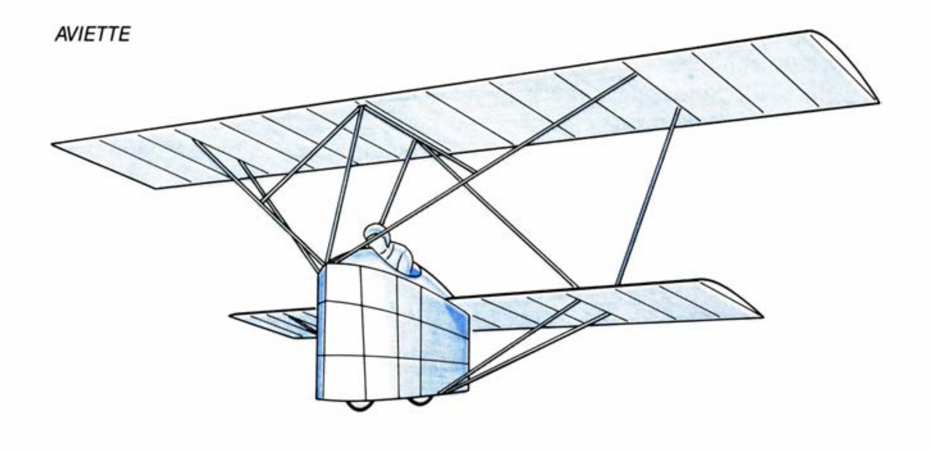


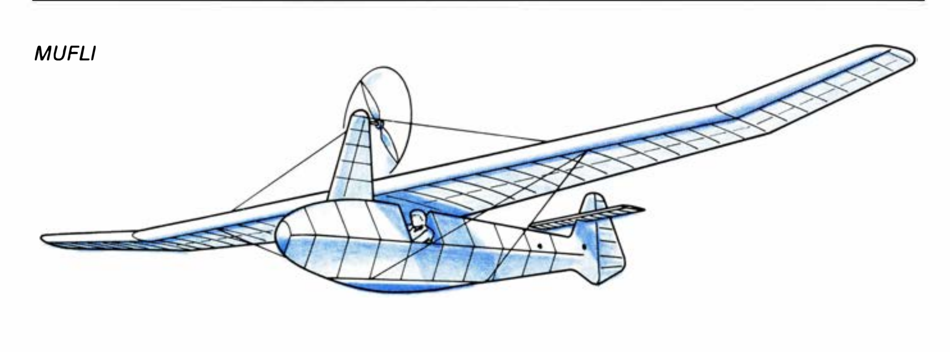
Scarabino, quien en el año 1984 ganó con esta nave el premio de 20.000 libras ofrecido por el británico Henry Kremer, al completar un trazado triangular de 1500 metros en menos de tres minutos. Las reglas de las competiciones de Kremer y sus sucesoras, patrocinadas por la Royal Aeronautical Society, estipulan que la nave debe volar a una altitud mínima de dos metros al comienzo y al final. La cinta naranja sujetada por dos miembros del personal de tierra está a dos metros de altura. La persona de la derecha también tiene un cronómetro, como el juez oficial situado en segundo plano (nombrado por la Royal Aeronautical Society). Cuando ganó el Kremer, el *Monarch B*, que incorpora un dispositivo electrónico para la hélice, voló a una velocidad de 21,5 millas por hora.

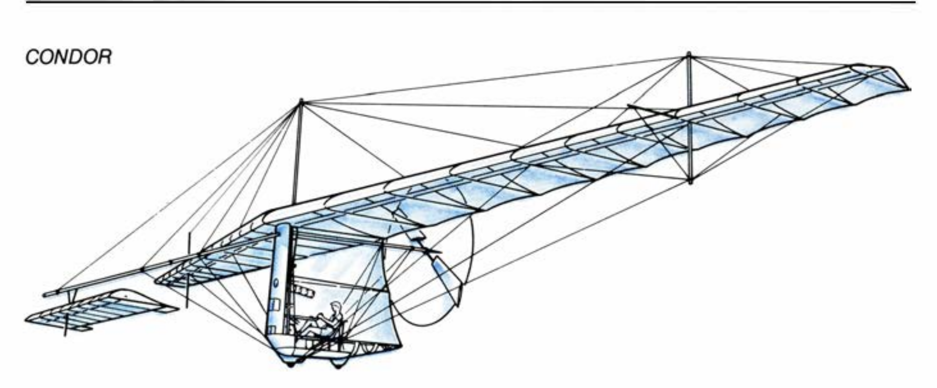
sido distintas, los ingenieros de las tres generaciones se las han tenido que ver con un problema común: cómo reducir la potencia requerida por la aeronave a una cantidad accesible para un ser humano. El segundo problema en importancia del vuelo de propulsión humana ha sido la estabilidad y el control.

La potencia generada por un ser humano difiere ampliamente según la edad de la persona, su entrenamiento y su afán. Un atleta bien preparado puede proporcionar hasta un kilowatt durante breves períodos de tiempo o unos cientos de watt durante varias horas. A la vista de los muchos estudios realizados, es sorprendente que no se disponga de conclusiones definitivas sobre factores básicos, tales como si es mejor para el piloto estar sentado verticalmente o recostado. En ausencia de datos fisiológicos fiables, la decisión del diseño se adopta generalmente por razones de aerodinámica, estructura o distribución de pesos.

La potencia requerida por una aeronave es el producto de sus resistencia aerodinámica al avance por su velocidad. Por tanto, si se construye una nave de baja resistencia se necesita poca potencia; volar despacio también es una forma de reducir la potencia requerida.







2. TECNOLOGIA AVANZADA de las aeronaves de propulsión humana en el *Aviette*, el *Mufli* y el *Gossamer Condor*. El *Aviette*, ganador de una competición patrocinada por Peugeot entre 1912 y 1922, era una bicicleta saltadora; no tenía fuente de propulsión tras abandonar el suelo, por lo que apenas planeaba unos metros. El *Mufli*, nave alemana de 1935, tenía una hélice de propulsión humana, pero el piloto sólo podía generar potencia suficiente para un largo planeo (712 metros el mayor). El *Condor* representa el tipo de nave de propulsión humana que puede volar una distancia indeterminada y es completamente controlable. En 1977 el *Condor* ganó la primera competición de Kremer volando una milla en un trayecto con figura de ocho.

Para alcanzar el equilibrio en vuelo, la sustentación (fuerza vertical) producida por las alas de una aeronave debe igualar el peso total del vehículo. La superficie de ala es la variable más útil para el diseñador. En teoría, con alas suficientemente grandes se pueden obtener velocidades de vuelo y necesidades de potencia tan bajas como se quiera. En la práctica, la superficie de ala está limitada por la rigidez estructural, peso, sensibilidad al viento y capacidad de los hangares.

A velocidades subsónicas, la resistencia tiene dos componentes de magnitud comparable. Se refiere la primera al rozamiento con el aire. Es más o menos proporcional al área frontal del aeroplano. La segunda componente es consecuencia inevitable de la generación de sustentación y se denomina resistencia inducida. La resistencia de rozamiento puede menguar si acortamos el área frontal y optamos por formas aerodinámicas eficientes. La resistencia inducida puede disminuirse, sobre todo, aumentando la envergadura y volando cerca del suelo. Factores teóricos y prácticos limitarán la cantidad de resistencia que se reduzca en cada paso.

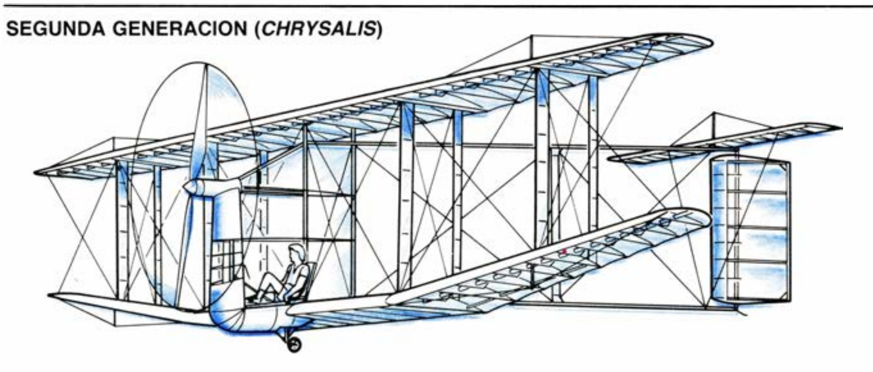
Los planeadores han representado siempre el ideal del diseño de baja resistencia. Era, pues, natural que la primera generación de aeronaves de propulsión humana se les pareciera. Los diseñadores buscaron principalmente reducir el peso del planeador en un factor de 10 a la vez que añadían una hélice, sin comprometer los principios aerodinámicos. Toda la estructura y el cableado eran interiores. Ahora sabemos que las pretensiones trascendían la capacidad de la técnica estructural entonces disponible. Las aeronaves resultaban pesadas y pequeñas, comparadas con las máquinas de la segunda generación. La baja resistencia y la relativamente alta velocidad de vuelo creaban unas necesidades de potencia que apenas si dejaban margen al piloto para maniobrar el vehículo.

En las naves de segunda generación se buscó ya conseguir velocidades bajas para reducir la potencia. Se abandonó la ventaja que suponía una baja resistencia de los planeadores en favor del cableado exterior. El incremento resultante en la resistencia se compensaba con un incremento sustancial de

PRIMERA GENERACION (*JUPITER*)

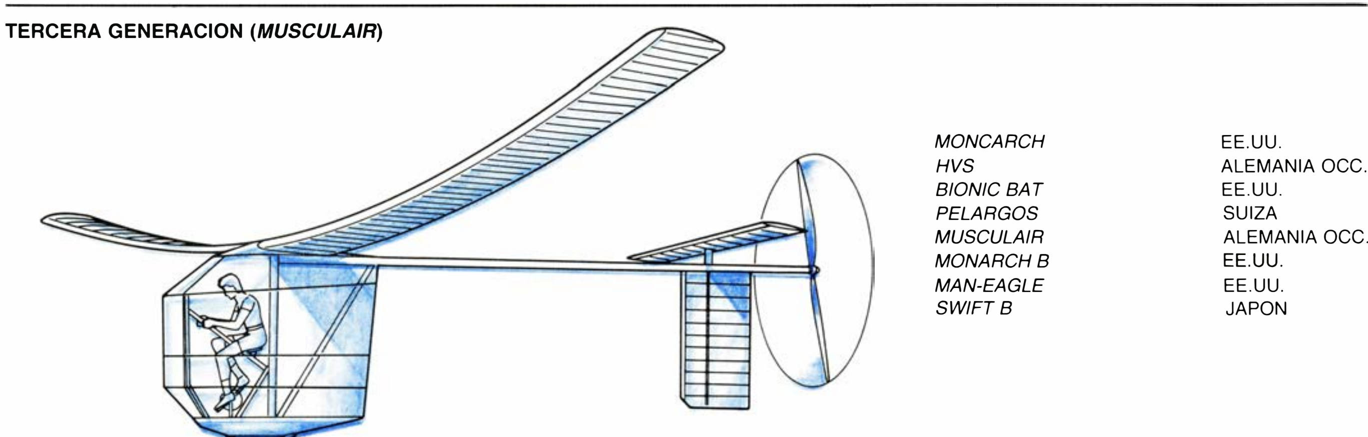
NOMBRE	ORIGEN
MUFLI	ALEMANIA
PEDALIANTE	ITALIA
SUMPAC	GRAN BRETAÑA
PUFFIN I	GRAN BRETAÑA
McAVOY	EE.UU.
VINE	SURAFRICA
MAYFLY	GRAN BRETAÑA
PUFFIN II	GRAN BRETAÑA
LINNET I	JAPON
RELUCTANT PHOENIX	GRAN BRETAÑA
LINNET II	JAPON
MALLIGA	AUSTRIA
SM-OX	JAPON
LINNET III	JAPON
LINNET IV	JAPON
MERCURY	GRAN BRETAÑA
OTTAWA	CANADA
WRIGHT	GRAN BRETAÑA
JUPITER	GRAN BRETAÑA
TOUCAN I	GRAN BRETAÑA
LIVERPUFFIN	GRAN BRETAÑA
EGRET I	JAPON
EGRET II	JAPON
EGRET III	JAPON
BURD I	EE.UU.
AVIETTE	FRANCIA
EGRET IV	JAPON
DEDAL III	POLONIA
TOUCAN II	GRAN BRETAÑA
STORK I	JAPON
BURD II	EE.UU.
BLIESNER	EE.UU.
OLYMPIAN ZB-1	EE.UU.
ICARUS	EE.UU.
SKYCYCLE	EE.UU.
STORK II	JAPON
NEWBURY MANFLIER	GRAN BRETAÑA
PHOENIX	GRAN BRETAÑA
PHILLIPS	GRAN BRETAÑA

SEGUNDA GENERACION (*CHRYSALIS*)



GOSSAMER CONDOR	EE.UU.
CHRYSALIS	EE.UU.
GOSSAMER ALBATROSS	EE.UU.
GOSSAMER PENGUIN	EE.UU.
MILAN '82	JAPON

TERCERA GENERACION (*MUSCULAIR*)



MONCARCH	EE.UU.
HVS	ALEMANIA OCC.
BIONIC BAT	EE.UU.
PELARGOS	SUIZA
MUSCULAIR	ALEMANIA OCC.
MONARCH B	EE.UU.
MAN-EAGLE	EE.UU.
SWIFT B	JAPON

3. AGRUPACION DE LAS AERONAVES en tres generaciones sucesivas. Refleja las principales diferencias en la tecnología del vuelo de propulsión humana. Las aeronaves de la primera generación tenían armazones interiores de madera; pesadas y frágiles, sólo volaban en línea recta. Las aeronaves de la segunda generación tenían estructura de tubos de aluminio y su cableado se componía de tirantes exteriores. Estas fueron las primeras naves completamente controlables. Las de la tercera generación son menores y más veloces. Los materiales actuales, pensemos en el grafito, permiten la construcción en voladizo.

superfice de ala y una gran reducción de peso. Con las bajas velocidades de vuelo resultantes (aproximadamente 16 kilómetros por hora) las necesidades de potencia descendieron.

Las máquinas de la segunda generación también incorporaron las primeras soluciones al problema de la estabilidad y el control. Este logro significaba que los diseñadores se habían enfrentado con éxito a diversos efectos que normalmente no son importantes en las aeronaves convencionales.

Citemos, por ejemplo, la aceleración. Para acelerar una aeronave (supongámosla realizando un giro inclinado) debe acelerarse también parte del aire que la rodea. Se dice entonces que la nave tiene una "masa aparente", además de la suya propia. En las aeronaves convencionales esta componente adicional es despreciable. En las aeronaves de propulsión humana adquiere especial importancia. En virtud de ello, las superficies de control convencionales no pueden generar las fuerzas necesarias para tratar adecuadamente la masa aparente, de manera que los diseñadores tuvieron que abordar el problema desde otro punto de vista.

Las aeronaves han de controlarse en tres ejes: guiñada, o giro respecto a un eje vertical, cabeceo y alabeo [*véase la figura 4*]. Generalmente, un timón de dirección vertical en la cola gobierna la guiñada, un timón de profundidad horizontal en la cola establece el cabeceo y los alerones horizontales de las alas determinan el alabeo o balance. Para iniciar un giro, el piloto balancea la nave por medio de los alerones. Esta acción inclina el vector de sustentación del ala, proporcionando la fuerza lateral necesaria para el giro. El timón de dirección se utiliza entonces para "coordinar" el giro, manteniendo el morro apuntado hacia la corriente. Los alerones controlan la velocidad de alabeo; se vuelven a dejar centrados una vez iniciado el giro y se usan otra vez para alabear la nave en sentido contrario cuando se quiere acabar el giro.

Cuando deflectamos los alerones, éstos imponen al ala un par que tiende a torsionarla a lo largo del eje de envergadura. El cambio resultante en el ángulo de ataque de los extremos de las alas (y por tanto en la sustentación) anula parcialmente el efecto de los propios alerones. Así pues, un control adecuado del alabeo exige del ala suficiente rigidez para resistir el par de torsión de los alerones.

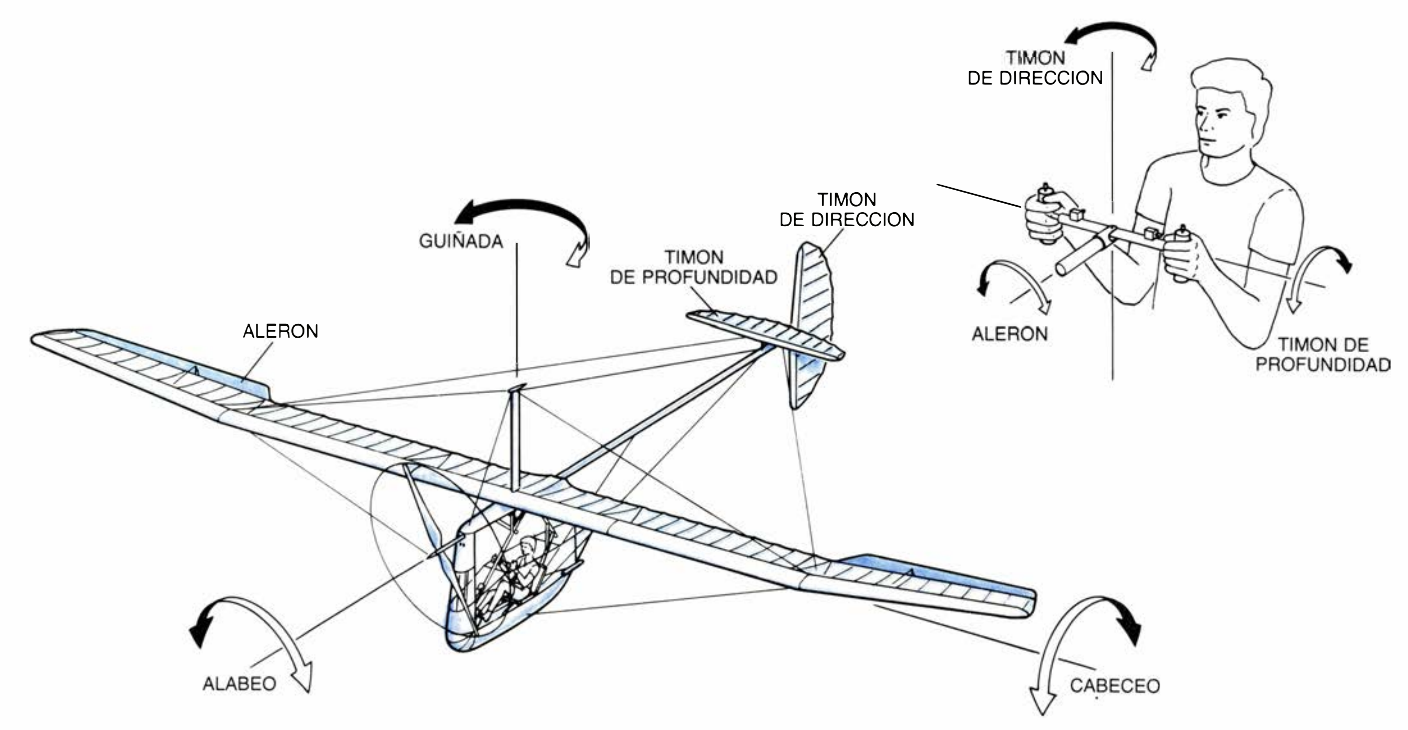
En las dos primeras generaciones de aeronaves de propulsión humana, la combinación de grandes masas aparentes y alas débiles a la torsión quitaba toda eficacia a los alerones. El problema se resolvió en el *Gossamer Condor* por medio de un pato (*canard*): una superficie de control montada en el fuselaje, por delante del ala. En el *Condor*, el pato se inclinaba produciendo una fuerza lateral similar a la generada por el timón de dirección, consiguiendo así la guiñada deseada. El movimiento de guiñada producía una mayor velocidad del aire y sustentación en el extremo del ala exterior y una menor velocidad del aire y sustentación en la interior. La diferencia de sustentación producía, pues, un alabeo en la nave.

Para evitar que la nave se inclinara demasiado, el piloto que volaba el *Condor* tenía que tirar del cableado exterior para torcer las alas, como lo hacían los hermanos Wright en su *Flyer*, en 1903. La maniobra aumentaba el ángulo de ataque (y, por tanto, la sustentación) del ala interior y disminuía el de la exterior. Esta acción posibilitaba giros controlados sostenidos.

Por ser menores las aeronaves de propulsión humana de la tercera generación, los efectos de masa aparente lo son también, y el ala puede adquirir bastante más rigidez. Los alerones han demostrado su utilidad en estas máquinas.

Volvamos ahora a los tres desarrollos tecnológicos cruciales para el éxito del vuelo de propulsión humana. Son los perfiles de alta sustentación, los sistemas de propulsión eficientes y las estructuras ligeras.

La principal superficie aerodinámica es el ala. Debido a que produce la mayor parte de la resistencia, la forma de su sección transversal (el perfil) ha de



4. CONTROL EN TRES EJES de una nave típica de la tercera generación. Se consigue con sólo las manos del piloto. (Sus pies pedalean para proporcionar la potencia de vuelo exigida.) El piloto controla el alabeo con los alerones, el cabeceo con el timón de profundidad y la guiñada con el timón de dirección.

alcanzar la mayor eficiencia posible. Una medida de la eficiencia de un perfil es la relación entre sustentación y resistencia (*S:R*). Otra medida de sus cualidades es el "parámetro de potencia", que es semejante al *S:R* pero concede más importancia a una elevada sustentación. Cuanto mayor es el parámetro de potencia, menor es la potencia necesaria para mantener el vuelo. Lo principal en las aeronaves de propulsión humana es la baja potencia; por tanto, un parámetro de potencia alto importa más que un gran *S:R*. Para obtener un parámetro de potencia alto, un perfil debe ser capaz de proporcionar una elevada sustentación sin inducir una resistencia excesiva.

El perfil debe tener también un momento de cabeceo pequeño, es decir, ha de tender a mantenerse en un mismo plano a lo largo del eje de vuelo. Un momento de cabeceo elevado genera los mismos pares sobre el eje de la envergadura que los alerones. (Esta es otra razón por la que las alas deben ser rígidas a la torsión.) Mayor rigidez que añade, inevitablemente, peso al ala. Además, un momento de cabeceo alto desestabiliza la aeronave y requiere mayores superficies de cola, lo que añade peso y resistencia.

Un factor que complica el diseño de aeronaves de propulsión humana es que funcionan en un régimen aerodinámico poco usual, competencia habitual de grandes aves y de aeromodelos. Este régimen está caracterizado formalmente por un número de Reynolds relativamente bajo, que es un factor de mérito adimensional que tiene en cuenta la velocidad, la densidad y la viscosidad del aire junto con la longitud del cuerpo alineado con el flujo. Los números de Reynolds de las aeronaves típicas oscilan entre dos y 20 millones. Se ha reunido abundante información sobre este tipo de vuelo desde la primera guerra mundial. Las aeronaves de propulsión humana funcionan con números de Reynolds inferiores al millón, una región de vuelo que no conocemos todavía bien.

El bajo número de Reynolds y la necesidad de elevada sustentación, baja resistencia y bajo momento de cabeceo han obligado a los ingenieros de aeronaves de propulsión humana a diseñar nuevos perfiles o adaptar los existentes. El problema estriba en ajustar la distribución de presiones en la superficie del perfil. Para entendernos, se puede decir que se emplean dos tipos de perfiles en las aeronaves de propulsión humana: de carga trasera y de carga frontal. Los términos reflejan el hecho de que la distribución de presiones en la parte superior e inferior del ala tiende a ser desigual, de suerte que la mayor parte de la carga viene soportada por la parte trasera del ala o por la frontal, según la elección del ingeniero proyectista.

Un perfil típico de carga trasera ofrece una alta relación de sustentación a resistencia en una gama de velocidades y ángulos de ataque bastante amplia. Funciona bien en planeadores, pero no tanto en las aeronaves de propulsión humana. Su principal desventaja es el elevado momento de cabeceo. Sin embargo, ésta y otras desventajas pierden importancia al disminuir el tamaño de la aeronave. La alemana *Musculair*, una nave muy lograda de tercera generación, utilizó un perfil de carga trasera.

Los perfiles de carga frontal ofrecen altas relaciones de sustentación a resistencia, así como los mayores parámetros de potencia, pero sólo en una gama relativamente estrecha de velocidades y ángulos de ataque. Aunque esta desventaja los hace inadecuados para los planeadores y la mayoría de aeroplanos, es ideal para aeronaves de propulsión humana, que forzosamente tienen una estrecha gama de velocidades en razón de su limitada potencia. Más aún, estos perfiles tienen bajos momentos de cabeceo, por lo que las alas con perfil de carga frontal pueden construirse con menos peso estructural que las alas con perfil de carga trasera.

La propulsión es otra área de las aeronaves de propulsión humana donde reviste interés conseguir un alto rendimiento. La hélice es, con mucho, el medio más eficaz para transformar la potencia mecánica generada por las piernas del piloto en un empuje suficiente para vencer la resistencia de la máquina. Cabe imaginar otros medios de propulsión para estas aeronaves, como alas batientes y chorros de aire comprimido, pero todavía no han tenido éxito.

Cualquier sistema propulsor que genera empuje (exceptuando los cohetes), toma aire a la velocidad de vuelo y lo expulsa por detrás a una velocidad mayor, en forma de chorro. En el caso de la hélice, el chorro es la corriente de aire que expulsa hacia atrás. El ala batiente empuja hacia atrás una masa amorfa de aire en cada golpe.

En todos los casos, el chorro se lleva una energía cinética que ha sido añadida por el sistema de propulsión y que no puede recuperarse: finalmente se disipa en forma de calor. Al aumentar la velocidad del chorro, la pérdida en energía malgastada crece más deprisa que lo ganado en empuje. Es lógico, por tanto, que para lograr altos rendimientos se desarrollen sistemas que tomen una gran masa de aire y le añadan sólo un pequeño incremento de velocidad. Este objetivo exige hélices de gran diámetro o alas batientes de gran envergadura. (El chorro de aire comprimido es, consecuentemente, de bajo rendimiento a las velocidades de las aeronaves de propulsión humana debido a su alta velocidad de chorro.)

Aunque, en teoría, el rendimiento del ala batiente puede llegar a ser muy alto, nunca se ha aplicado con éxito a ninguna nave tripulada. Para obtener buen rendimiento, el ala debe doblarse en una dirección a lo largo del eje de su envergadura en el recorrido descendente y en la otra dirección en el ascendente. Las aves realizan muy bien esta maniobra, pero en una máquina la combinación entre el movimiento de batir las alas y doblarlas plantea serios problemas mecánicos y estructurales que empeoran al aumentar el tamaño de la nave. De ahí que la hélice sea generalmente el único sistema de propulsión práctico en las aeronaves de propulsión humana.

El ideal de hélice de gran tamaño se enfrenta a varios problemas en las aeronaves de propulsión humana. Una hélice grande añade peso, algo que el diseñador trata de evitar. A partir de cierto tamaño existe la posibilidad de que los extremos de la hélice golpeen el suelo cuando la nave despega o aterriza. Así que el diseñador no puede alcanzar el máximo rendimiento con sólo aumentar el diámetro de la hélice. En lugar de eso, debe intentar reducir la pérdida de eficiencia, debida a la energía cinética que tiene la corriente de aire producida por la hélice, estudiando cuidadosamente la distribución de carga sobre la pala. El rozamiento del aire en las palas influye también en el diseño de la hélice. Un diámetro grande y un diseño óptimo permiten que las hélices de las aeronaves de propulsión humana obtengan rendimientos próximos al 90 por ciento.

La tecnología estructural es la característica de las aeronaves de propulsión humana que más ha cambiado desde los vehículos de la primera generación. En aquellas aeronaves, complejos armazones construidos principalmente de

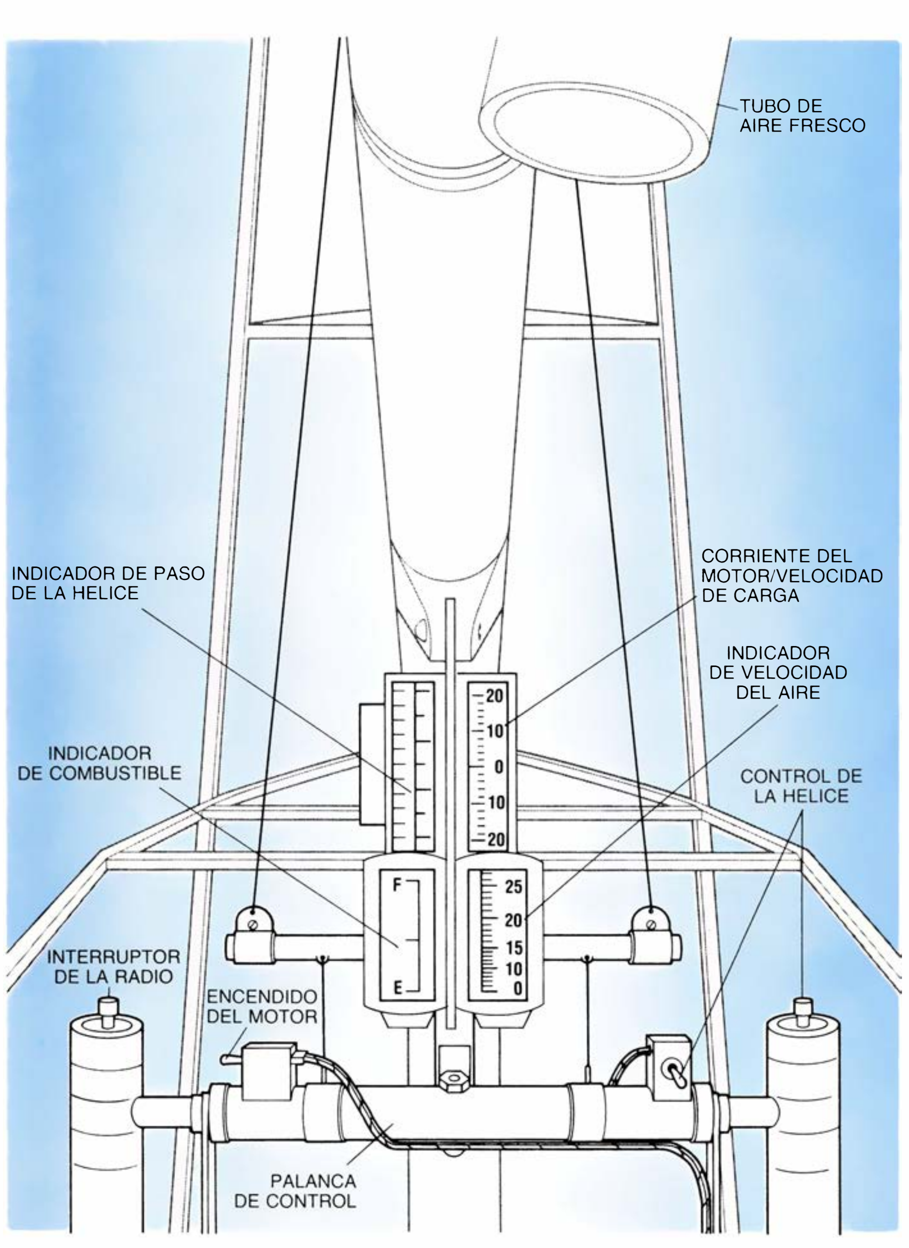
madera decidían la forma y resistencia del ingenio. El armazón es una estructura eficiente: da altas relaciones de esfuerzo a peso y de rigidez a peso. La madera resulta fácil de obtener, fácil de trabajar y relativamente barata. Más aún, la mayoría de los que fabrican aeronaves de propulsión humana son o eran entusiastas del aeromodelismo y, por tanto, expertos en trabajar la madera.

Por otra parte, el armazón de madera presenta varios inconvenientes. Consta de tantas juntas y piezas aisladas que la construcción da mucho trabajo. Arreglar una pieza rota es difícil. Además, si falla una parte del armazón, las piezas cercanas quedan sometidas a un esfuerzo para el que no están preparadas, lo que pone en peligro toda la estructura. Por estas razones se abandonó el armazón de madera en las aeronaves de la segunda generación. Los diseñadores se inclinaron por una estructura principal de tubos de aluminio de gran diámetro y pared delgada; los cables se encargaban de las ligaduras exteriores.

Los tubos se proyectaban ante todo para resistir la compresión. Los cables exteriores soportaban todas las cargas principales de flexión y torsión. (A bajas velocidades de vuelo, la resistencia causada por los cables está más que compensada por la disminución de peso.) La ventaja de una estructura así es que tiene una alta relación de firmeza a peso y proporciona una excelente rigidez. La ausencia de armazones de madera facilitaba también la reparación de las naves de la segunda generación, mucho más que sus predecesoras.

Debido a que las naves de la tercera generación son menores y se dispone de materiales más fuertes, como el grafito y el grafito-epoxy, los ingenieros pudieron volver a estructuras en voladizo, que eliminan los cables exteriores. También la película de mylar como cubierta ha mejorado la estructura.

La combinación de baja velocidad, baja altitud y potencia limitada ha convertido en un reto pilotar una aeronave de propulsión humana. Pero está al alcance casi de cualquiera. Las naves de uso general, pensemos en *Condor* y *Chrysalis*, han sido voladas por hombres y mujeres de edades entre los dieciocho y los sesenta.



5. SISTEMAS DE CONTROL del *Monarch B*. Los ilustramos tal cual los ve el piloto. El indicador de corriente/carga muestra la velocidad a la que se carga o descarga la batería de la nave; la carga tiene lugar antes del vuelo, cuando el piloto pedalea un generador para almacenar potencia. Durante el vuelo, puede usar la potencia almacenada por medio de un interruptor, que enciende un motor eléctrico. El indicador de combustible señala la cantidad de carga de la batería. El indicador de paso de la hélice refleja el ángulo de las palas, que el piloto puede controlar por medio del interruptor de control de hélice. Apretando el interruptor de la radio, el piloto puede hablar con el personal de tierra y oir en todo momento. La palanca de control del *Monarch* lleva un mensaje de aliento al esforzado piloto: "Tú tienes grandes facultades físicas".

Generalmente, el piloto empieza por mudarse de ropa. Un equipo de deporte y un casco de bicicleta constituyen el atuendo apropiado: el peso extra exige potencia extra, y la luz del sol sobre la cubierta transparente hace incómodo el puesto de pilotaje cuando la nave no se mueve.

Es difícil entrar en la aeronave sin dañarla. Hay pocos lugares lo bastante sólidos para soportar el peso de uno, por lo que el piloto utiliza normalmente plataformas de acceso y es ayudado por miembros del personal de tierra. Tras repasar una hoja de comprobación antes del vuelo, el piloto da la señal a los miembros del equipo que sujetan los extremos de las alas y empieza a pedalear.

Se despega con sorprendente suavidad. La mayoría de los que vuelan por primera vez ignoran que la nave está en el aire hasta que oyen los vítores del personal de tierra. La cabina del piloto es más ruidosa de lo que cabría esperar, debido al chirrido de la cadena de la bicicleta y al cíclico zumbido de la hélice al pasar las palas por la carena.

La principal tarea del piloto en el aire es concentrarse en mantener una altitud y una velocidad estables. Si se ha equilibrado correctamente la nave para su peso, sólo se requerirán pequeños ajustes del timón de dirección. Para elevarse, el piloto pedalea más fuerte; para bajar, basta aminorar la velocidad de pedaleo.

La aeronave de propulsión humana reacciona muy despacio a los controles, por lo que los pilotos inexpertos pueden llegar a perder el dominio de la misma. Más confusa resulta todavía la tendencia de la aeronave a responder de forma diferente según los distintos ejes. La respuesta al cabeceo es bastante rápida; la del alabeo desesperadamente lenta. Realizar giros siguiendo un trazado específico, como se exige en las competiciones de velocidad, requiere una cuidadosa coordinación y mucha práctica.

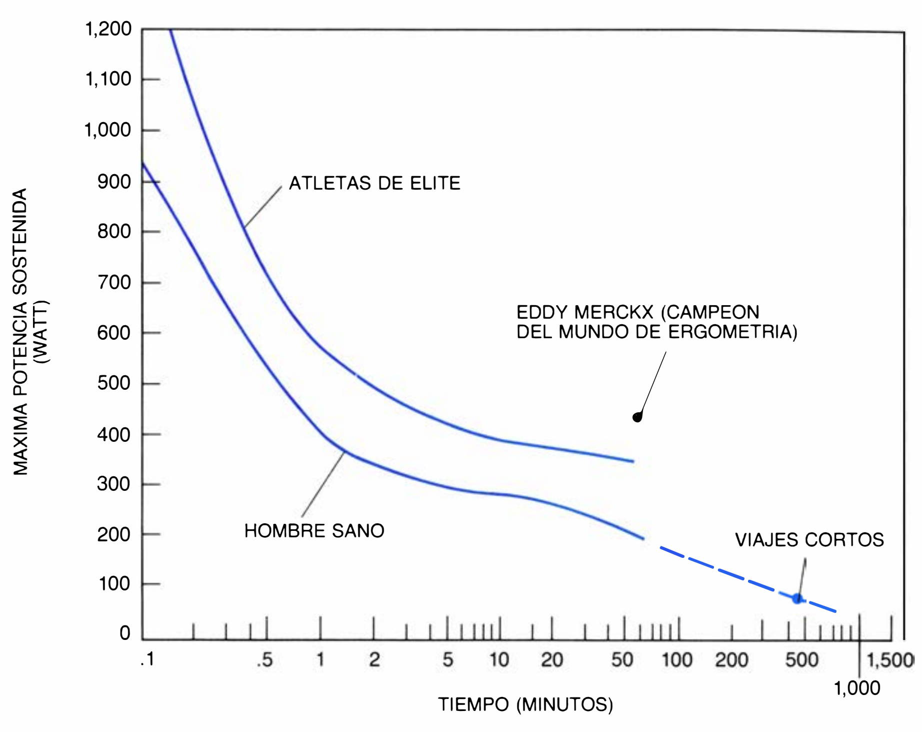
Los dos principales peligros en el vuelo son la pérdida y las ráfagas de viento. La nave entra en pérdida cuando el flujo de aire no se corresponde con la superficie del ala. Normalmente, esa discrepancia ocurre cuando el piloto ha dejado que la velocidad de la nave descienda demasiado.

Las velocidades de vuelo necesariamente bajas hacen de las ráfagas un problema especial. Por ser crucial la relación entre la velocidad del viento y la de la aeronave, una ráfaga de sólo ocho kilómetros por hora equivale a una de 50 kilómetros por hora o más para un pequeño avión convencional. Chocando de frente, una ráfaga así puede producir una sobrecarga y romper las alas; si viene por detrás puede producir la pérdida. Las ráfagas pueden variar también la trayectoria de vuelo y la altitud del aparato. Afortunadamente, las bajas velocidades y altitudes se combinan para que las aeronaves de propulsión humana sean bastante seguras: accidentes que destruyen la estructura apenas si producen en el piloto más que cortes y magulladuras.

Para aterrizar, el piloto alinea la nave con la pista y reduce la velocidad de pedaleo. El vehículo planea despacio y toma tierra suavemente.

El vuelo de propulsión humana se ha desarrollado sin mayores pretensiones, por el puro estímulo de las distintas competiciones. Sin embargo, puede esperarse que las tecnologías desarrolladas tengan aplicaciones prácticas en tres campos por lo menos: el propio vuelo de propulsión humana, aeronaves ultraligeras y distintas tareas de reconocimiento y observación.

En el vuelo de propulsión humana, las competiciones de velocidad continuarán, tal como están establecidas ahora, hasta que se ganen 11 premios más de 5000 libras. Por supuesto, las aeronaves serán más rápidas, alcanzando quizá velocidades de 85 kilómetros por hora. Pero esa velocidad no será fácil de lograr, ni siquiera con sistemas de almacenamiento de energía de gran rendimiento. Queda por ver si los premios propuestos bastarán para provocar las necesarias inversiones en dinero e ideas.



6. VARIACION DE LA POTENCIA HUMANA según la edad, condición y voluntad de la persona. La escala se indica en este gráfico. La línea de "viajes turísticos", o cortos, muestra valores deducidos de las carreras ciclistas campo a través. Un kilowatt (1000 watt) es equivalente a 1,3 caballos de vapor. Los datos reseñados proceden de *Bycicling Science*, del que son autores Frank Rowland Whitt y David Gordon Wilson.

La Royal Aeronautical Society está examinando la posibilidad de organizar nuevas competiciones. Probablemente, tendrían como objetivo que las aeronaves de propulsión humana fueran más prácticas y robustas. En el otro extremo del espectro, ya es posible (desde nuestro punto de vista) construir una aeronave grande y de baja potencia capaz de convertir en realidad la leyenda de Dédalo: volar los 96 kilómetros que separan a Creta de la Grecia continental a unos 22 kilómetros por hora.

La relación entre las aeronaves de propulsión humana y los ultraligeros (impulsados por pequeños motores de gasolina) se va estrechando: las primeras, diseñadas últimamente con vistas a la velocidad y la utilidad, se parecen cada vez más a los últimos. Las aeronaves de propulsión humana de la tercera generación emplean a velocidad de crucero unos 0,5 caballos de vapor, potencia que puede aumentarse con facilidad hasta dos caballos de vapor. Los motores de las aeronaves ultraligeras de hoy proporcionan de 30 a 50 caballos de vapor. Cabe esperar que los avances que registre la tecnología de las aeronaves de propulsión humana reduzcan esta diferencia para que puedan desarrollar algunas de las tareas que ahora exigen ultraligeros y que éstos puedan funcionar a potencias inferiores.

La última aplicación tiene relación con las operaciones a gran altitud. Se está considerando ahora la posibilidad de usar naves de gran altitud capaces de vuelos prolongados como plataformas, sin tripulación, para reconocimiento, transmisión de comunicaciones y trabajos de muestreo de la estratosfera. Una nave de gran altitud opera a los bajos números de Reynolds característicos de las aeronaves de propulsión humana. Por tanto, la tecnología desarrollada para elevar la resistencia estructural y reducir el peso de las aeronaves de propulsión humana beneficiará también a los vehículos de gran altitud.

Con el tiempo, estas técnicas podrían encontrar aplicación en el espacio. Piénsese en la atmósfera de Marte; a pesar de ser mucho menos densa que la terrestre, podría permitir el vuelo alado a números de Reynolds similares a los de las aeronaves de propulsión humana. Un vehículo alado, no tripulado (una aeronave análoga al *Lunar Rover*), constituiría una buena plataforma desde la cual examinar el terreno y tomar muestras de la atmósfera de Marte.

Tarjetas inteligentes

Las tarjetas provistas de microcircuitos son más seguras y versátiles que las tarjetas de crédito. Para funcionar en tan especial ambiente, las pastillas microelectrónicas han de cumplir muy estrictas exigencias

Robert McIvor

Tal vez dentro de un año el lector lleve ya en un departamento de la billetera gran parte de la potencia de cálculo de un ordenador personal. Residirá en un dispositivo llamado "tarjeta inteligente", y consistirá en una o más pastillas microelectrónicas ("chips") montadas en un rectángulo de plástico del tamaño de una tarjeta de crédito.

Ya se están produciendo en Francia tarjetas inteligentes cuya capacidad de cómputo es sólo ligeramente inferior a la de un ordenador personal; tarjetas que se emplean en varias aplicaciones. En el pasado mes de septiembre, se entregaron 50.000 tarjetas inteligentes a usuarios americanos (en Washington D.C. y en Palm Springs, Florida), formando parte de las pruebas de campo a que las está sometiendo Mastercard. Es posible que se acabe por producir muchos tipos de tales tarjetas, de distintos niveles de refinamiento y prestaciones. Las más perfeccionadas podrían portar pequeños visores de cristal líquido y disponer de capacidad para cifrar y descifrar cualquier diálogo con un dispositivo externo; estarían alimentadas por fotocélulas. La tarjeta menos refinada podría consistir en poco más que un procesador sencillo y una pequeña memoria, y podría servir, por ejemplo, como "tarjeta de cargo" para conferencias interurbanas a través de teléfonos públicos o como historial clínico personal. No es inconcebible que a la vuelta de algunos años las tarjetas inteligentes almacenen versiones digitalizadas de la firma de su titular, sus impresiones retinianas o sus huellas dactilares; se podrían utilizar tarjetas inteligentes debidamente individualizadas con la función de llaves de alta seguridad, que autorizasen el acceso a redes telefónicas, bancos de datos de uso colectivo y edificios seguros.

En la que seguramente sería su aplicación más natural, las tarjetas inteligentes podrían sustituir a las actuales tarjetas de crédito y a las utilizadas en los cajeros automáticos de los bancos. Una transacción típica con tarjeta inteligente se desarrollaría como sigue: el usuario insertaría la tarjeta en una "terminal de punto de venta", que es una máquina registradora especial, apta para lectura de tarjetas. La terminal le suministra a la tarjeta la energía eléctrica necesaria, y se comunica con los microcircuitos de ésta a través de ocho contactos metálicos situados en la superficie de la tarjeta. Se le solicita al usuario que facilite a la máquina una contraseña, que ha de escribir en un teclado; la máquina comprueba que la tarjeta sea válida y que el usuario sea el legítimo. Una vez materializada la operación de venta, el importe de la transacción se registra en la memoria de la tarjeta, siéndole acreditado al vendedor y deducido del saldo de crédito del titular; saldo que queda también almacenado en la memoria. El titular puede después reponer fondos en un cajero automático.

Las tarjetas de crédito y de cajeros automáticos, que las tarjetas inteligentes podrían llegar a sustituir, se valen para almacenar información de un sistema de banda magnética y de grabaciones en relieve. Las tarjetas de bandas magnéticas son totalmente pasivas (reservando su papel al de meros soportes para almacenamiento de información) y son vulnerables, ya que son susceptibles de muchas formas de abuso y falsificación. El tipo de abuso más sencillo consiste en gastar de más. Puesto que en la mayoría de las compras no se informa inmediatamente a la entidad que emitió la tarjeta, el cliente puede gastar mucho más del límite crediticio concedido, ya sea haciendo un gran número de transacciones de importes relativamente pequeños o bien una gran transacción no refrendada.

Los métodos de falsificación suelen ser mucho más refinados. En este sentido, se dispone de máquinas que copian la información almacenada en la banda magnética de una tarjeta de crédito sobre una tarjeta en blanco, o una imitación, al tiempo que sacan una impresión en papel carbón del huecograbado de la tarjeta. Estos dispositivos se parecen a las máquinas que utilizan los almacenes y los restaurantes para confeccionar recibos y para registrar las compras. Otro tipo de fraudes se basa en los números de identificación personal, o contraseñas del usuario. Cuando se utiliza una tarjeta magnética en un cajero automático, lo primero que éste hace es solicitar del usuario una contraseña. Seguidamente, lee en la tarjeta la clave de identificación correcta, y la compara con la contraseña recibida. Así pues, en algún momento del desarrollo de la transacción, la contraseña correcta ha de ser llevada a la memoria de trabajo del cajero automático. Si alguno de los "piratas informáticos" (un "hacker") da con un procedimiento para supervisar esa memoria, podrá hacerse con la contraseña del titular legítimo de la tarjeta.

Las tarjetas inteligentes tienen dos propiedades esenciales, que las hacen invulnerables a todos esos tipos de fraudes y abusos. Ante todo, una tarjeta inteligente tiene una memoria programable, no volátil, del tipo "de lectura únicamente", es decir, una memoria en la que se puede alojar información después de haberse emitido la tarjeta, y que recordará tal información aun estando desconectada de una fuente de alimentación. Esta memoria puede recordar el importe de cada transacción y el valor total de lo gastado, garantizando así que el usuario no rebase un límite preestablecido.

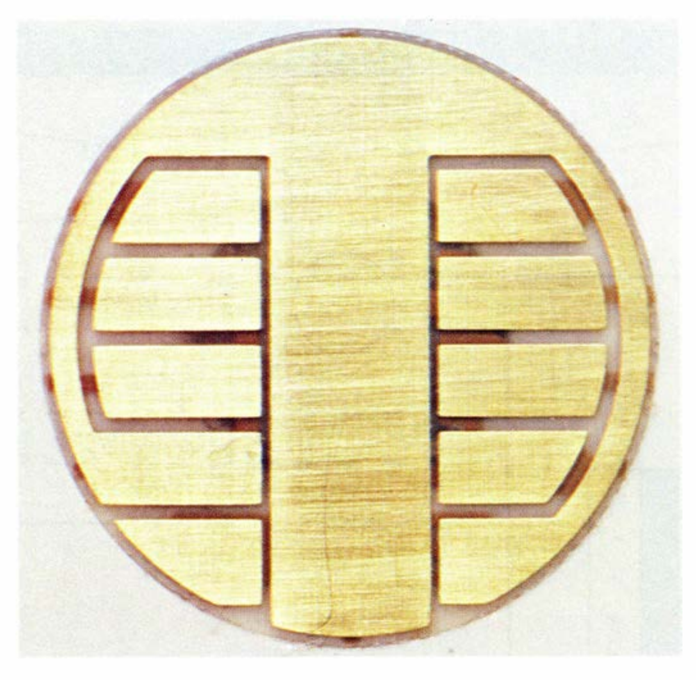
En segundo lugar, cada tarjeta inteligente contiene su propia unidad central de procesamiento (que es, en esencia, un pequeño ordenador), encargada de controlar todas las interacciones entre la memoria y los diversos disposi-

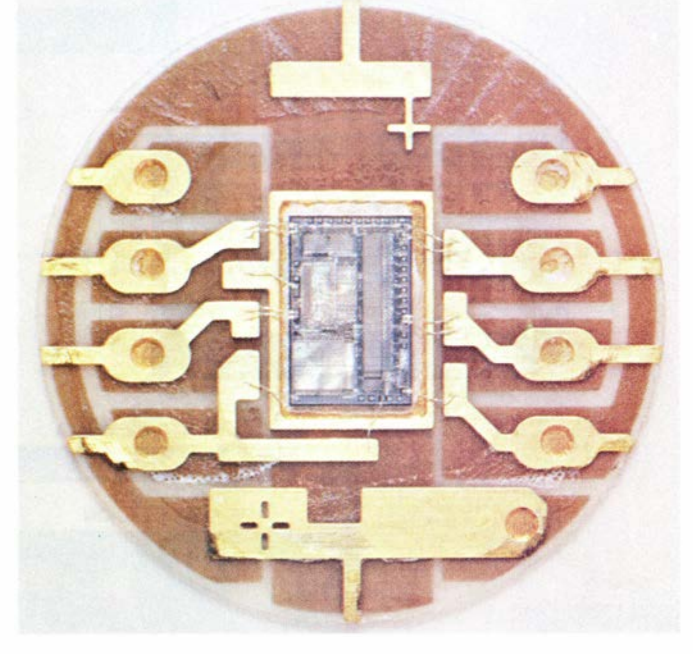
tivos externos que se precisan para leer la tarjeta e introducir datos en ella. La unidad central de procesamiento y la arquitectura de la memoria pueden construirse de modo tal que ciertas partes de la memoria de la tarjeta sean física o lógicamente inaccesibles para todo el mundo, exceptuada la entidad emisora de la tarjeta, pues la unidad central de procesamiento no obedecerá instrucciones de terceras personas que le ordenen leer o alterar dichas porciones de la memoria.

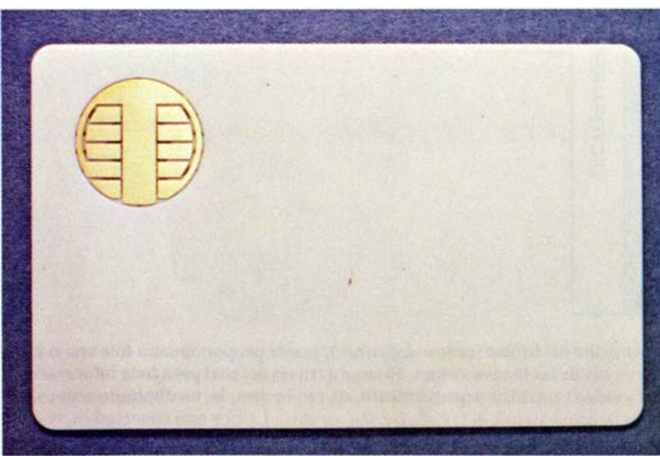
Merced a la unidad central de procesamiento de que la tarjeta va provista, ésta puede examinar cualquier contraseña que le sea presentada y compararla con la clave interna correcta, que está alojada en una posición secreta de la memoria de la tarjeta. La tarjeta no tiene en ningún momento que revelar la contraseña correcta a nada ni nadie ajeno al sistema. Ni siquiera la propia compañía emisora tiene por qué conocer tal contraseña. Cuando la tarjeta se emite, en el momento de serle entregada al titular, éste puede programar directamente en ella la contraseña mediante una máquina de lectura y grabación de tarjetas. Una vez introducida la contraseña, y verificada (al titular se le pide que introduzca su contraseña dos o tres veces, para garantizar que no se ha introducido incorrectamente), la tarjeta almacena la clave en la "zona secreta" de la memoria.

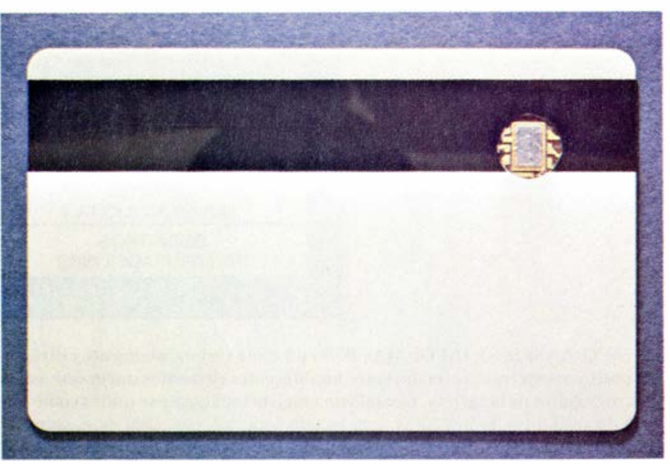
La zona secreta no sólo almacena la contraseña del titular, sino también su saldo bancario, el número de serie de la tarjeta y una serie de letras y números, elegidos por la entidad emisora, que pueden servir para verificar la legitimidad de la tarjeta. Otra zona de la memoria (memoria que es programable, pero solamente de lectura), llamada zona abierta, podría registrar el nombre del titular, su dirección, número de teléfono y número de cuenta bancaria. La zona abierta puede leerla cualquier máquina de lectura de tarjetas, pero no puede modificarse, pues la unidad central de procesamiento no obedecerá instrucciones que le ordenen alterar la información registrada en esta zona.

Cada vez que se utiliza la tarjeta para hacer una compra, las informaciones tales como el importe, el nombre y dirección del establecimiento comercial y la fecha quedan registradas en otra zona de la memoria, llamada zona de





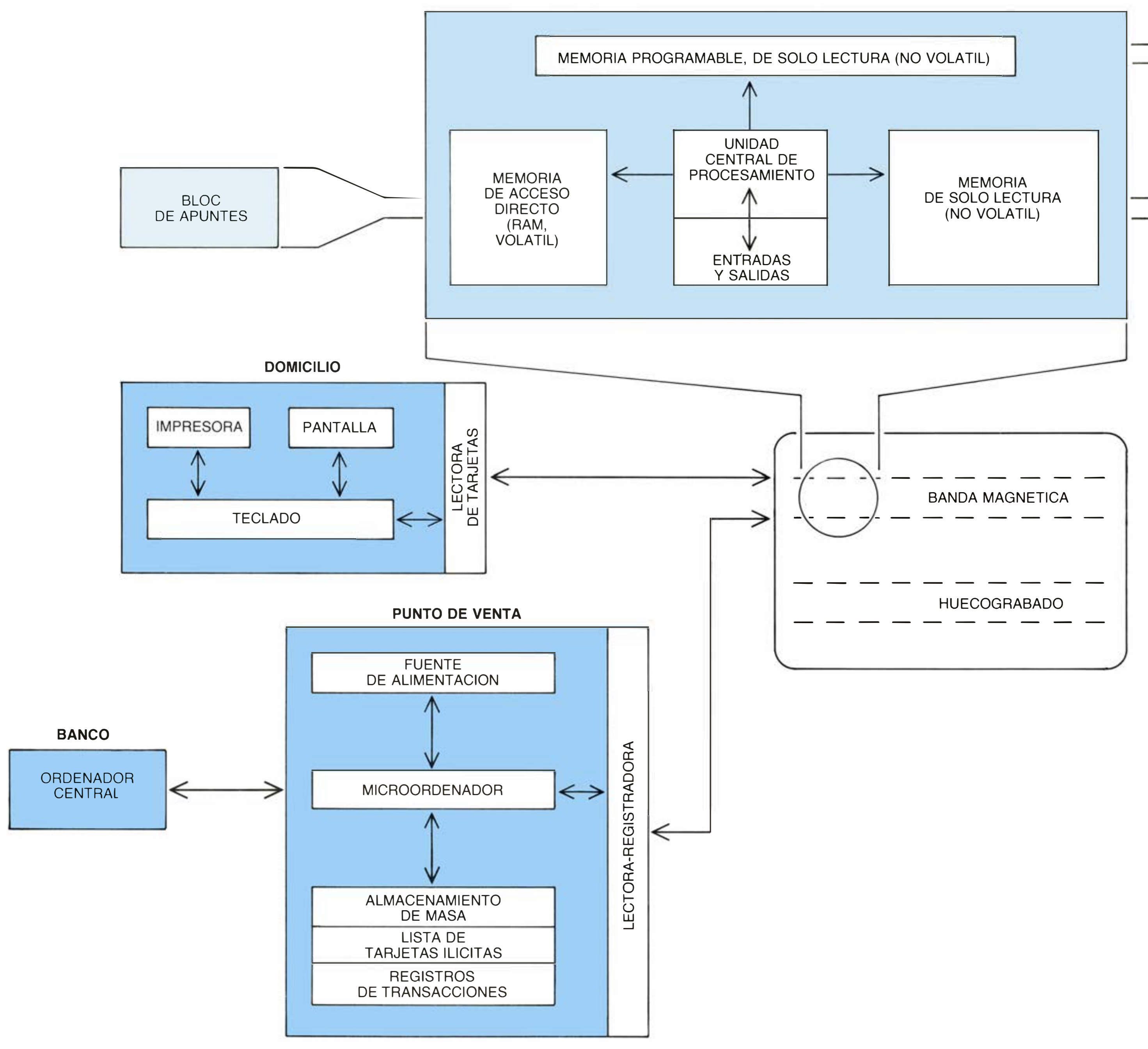




1. TARJETAS INTELIGENTES son tarjetas de plástico provistas de pastillas ("chips") microelectrónicas; se las puede usar como tarjetas de crédito, llaves u ordenadores portátiles. Así, la pastilla (*arriba, a la derecha*), que alberga un microprocesador y diversas clases de bancos de memoria, puede almacenar e ir actualizando el saldo de la cuenta de crédito del titular de la tarjeta y llevar un registro completo de todas las transacciones realizadas con la tarjeta. El microcircuito se comunica con la maquinaria exterior a través de ocho contactos metálicos (*arriba, a la izquierda*) situados en el anverso de la tarjeta (*abajo, a la izquierda*). En la fotografía del ángulo inferior derecho se ha retirado de la tarjeta parte de la banda magnética, con el fin de mostrar un posible emplazamiento de la pastilla. Esta podría también haberse situado justamente encima del eje longitudinal de la tarjeta. Composición icónica realizada por James Kilkelly.)

trabajo. Para registrar información en esta zona es preciso que concurran ciertas circunstancias (por ejemplo, que la tarjeta se encuentre en una caja registradora adecuada), y solamente se puede leer o escribir en ella con autorización de su titular. Por su parte, éste puede adquirir una máquina personal de lectura de tarjetas, la cual, conectada a un ordenador doméstico, a un televisor o a una impresora, muestra una relación completa de todas las compras efectuadas con la tarjeta.

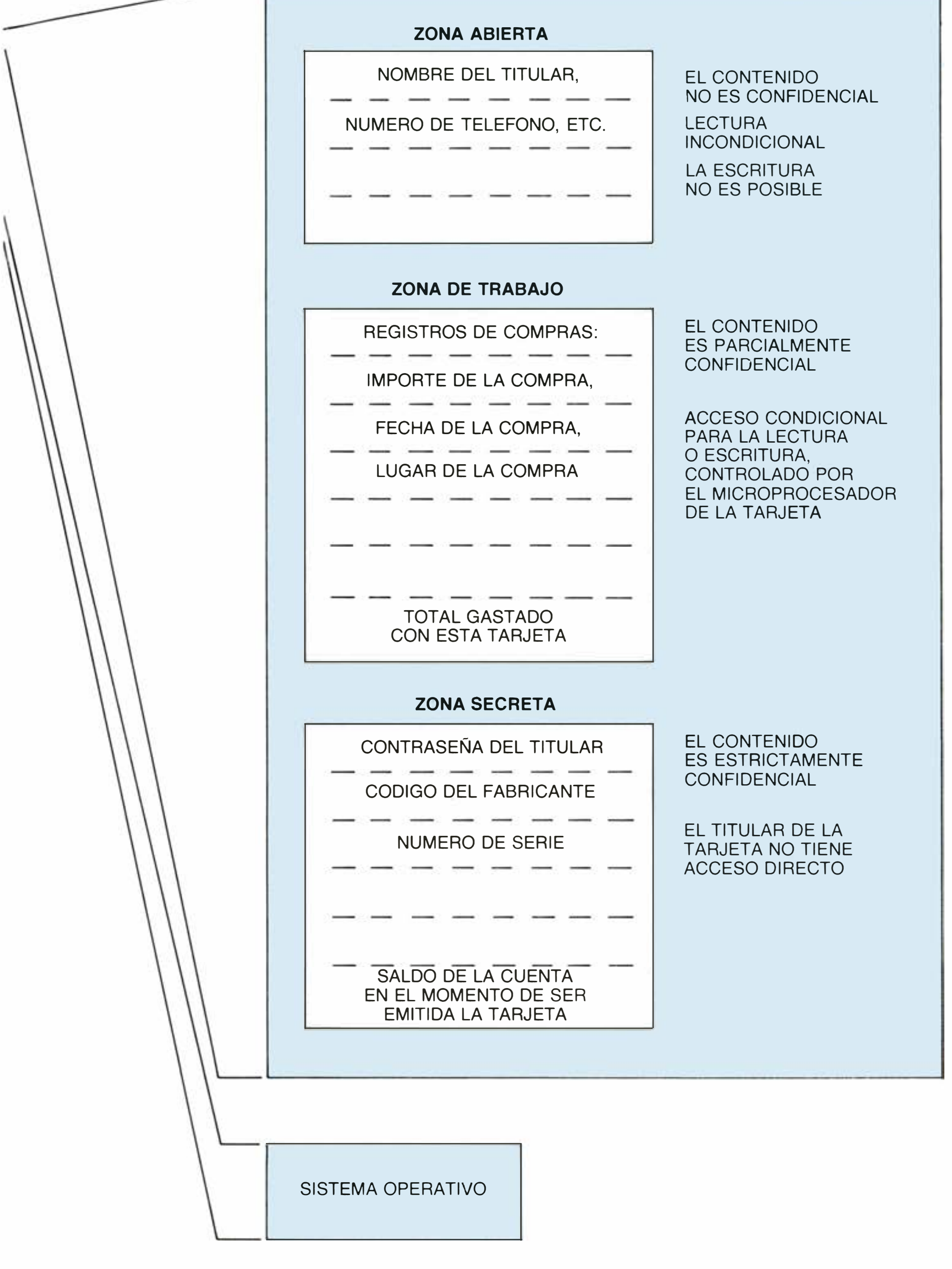
Al diseñar una tarjeta inteligente, el ingeniero se tropieza con una serie de condicionantes y restricciones especialmente graves. La primera de todas, que la tecnología existente, que se vale de tarjetas grabadas en relieve y de bandas magnéticas, limita seriamente el número de posibles lugares donde instalar la pastilla ("chip"). A causa del gran número de dispositivos que ya hay en servicio para la lectura de tarjetas en relieve y tarjetas magnéticas, no es posible cambiar ni la ubi-



2. SE TRANSFIERE INFORMACION entre una tarjeta inteligente y diversos dispositivos externos, así como entre los diferentes elementos integrados en los microcircuitos de la tarjeta. Cuando una tarjeta inteligente se utiliza como tarjeta de crédito, se la inserta en una lectora especial, instalada en el punto de venta, un comercio o un restaurante, pongamos por caso (*abajo, izquierda*). La máquina de lectura está conectada a un microordenador con acceso a una memoria que contiene una relación de tarjetas de crédito robadas y un registro completo de las transacciones realizadas por el comercio. La máquina emplazada en el punto de venta se comunica periódicamente con el ordenador de control de la entidad emisora de la tarjeta (*a la izquierda*) tanto para dar cuenta de las transacciones efectuadas como para actualizar la lista de tarjetas robadas de que el comercio tiene noticia. Una máquina de lectura especial, situada en el domicilio del titular (*centro, izquierda*), puede proporcionar a éste una relación impresa de las transacciones. El nexo a través del cual pasa toda información es la unidad central de procesamiento, o UCP (en esencia, un diminuto ordenador), que se encuentra en la tarjeta (*arriba, centro*). La UCP está conectada a tres tipos de memoria. El contenido de la memoria "de sólo lectura" (*a la derecha*) está determinado por el fabricante de la tarjeta, y no puede ser alterado; es una memoria no volátil (la información almacenada en ella permanece allí aun cuando la tarjeta esté desconectada de una fuente de alimentación). La memoria de sólo lectura (ROM) contiene el sistema operativo de la tarjeta, es decir, la secuencia de pasos que habrá de dar la unidad central de procesamiento cuando reciba una orden cursada por un dispositivo externo. Otra memoria, la de acceso directo (RAM) es volátil, es decir, incapaz de conservar la información una vez

cación de la banda ni la del huecograbado. Así pues, no cabe alojar la pastilla en la porción inferior de la tarjeta, donde se encuentra la grabación en relieve. Por otra parte, si se instalase en la mitad superior, podría interferir con la zona ocupada por el logotipo de la entidad emisora o, peor, con el funcionamiento de la banda magnética.

Hemos de considerar ahora dos posibles emplazamientos, que son los tenidos por menos inadecuados entre los aún disponibles. Si fuera posible instalar el microcircuito sin que afectase a la banda magnética, la pastilla podría alojarse en el ángulo superior izquierdo de la tarjeta. Alternativamente, podríamos situarlo ligeramente por encima del eje longitudinal, próximo a su borde izquierdo. Desafortunadamente, es ésta una región donde los esfuerzos mecánicos son máximos cuando la tarjeta resulta doblada en torno a su eje longitudinal. La finalidad de los ensayos de campo de Mastercard, en los cuales la mitad de las tarjetas llevan microcircuitos en una de estas posiciones, y la otra mitad en la otra, consiste, en parte, en determinar si la situación de la pastilla en la tarjeta afecta a la fiabilidad de ésta. La ubicación que resulte más aceptable se establecerá con carácter internacional por la Organización Internacional de Normalización (OIN).



desconectada de la fuente de alimentación. Se la utiliza como libreta de apuntes de la unidad central de procesamiento, es decir, como lugar donde almacenar temporalmente los resultados intermedios de los cálculos. La memoria de tercer tipo, de sólo lectura, pero programable (PROM), es la que confiere a la tarjeta casi toda su utilidad y versatilidad. Esta memoria programable, de sólo lectura, está dividida en tres zonas (*a la derecha*). Una de ellas, la zona secreta, únicamente resulta accesible para la entidad emisora de la tarjeta, pues la UCP no obedecerá instrucciones de nadie más que le ordene leer o modificar las informaciones de aquel sector. La zona secreta recoge informaciones tales como el tope crediticio del titular y su clave. Otras informaciones, como el nombre del titular, están almacenadas en una zona abierta. La zona abierta puede ser consultada por cualquier máquina de lectura de tarjetas, si bien las informaciones que contiene no pueden alterarse. La tercera zona, de trabajo, contiene un registro completo de las compras efectuadas con la tarjeta: importe de la compra, lugar y fecha de la misma, sin olvidar el monto del total gastado en todas las adquisiciones. La lectura o escritura de la zona de trabajo únicamente resulta posible bajo ciertas condiciones.

Hay otra cuestión que está decidida en gran medida por el medio en que han de desenvolverse los microcircuitos de las tarjetas inteligentes; a saber: ¿deben integrarse en un único microcircuito todos los procesadores y memorias, o bien deberían las tarjetas llevar cada una dos pastillas? En las tarjetas de dos pastillas ("chips"), una de ellas albergaría la unidad central de procesamiento, una memoria de trabajo, y poco más; la otra estaría formada casi totalmente por bancos de memoria.

Están disponibles ya, y pródigamente, pastillas ("chips") que contienen un microprocesador y una memoria de trabajo; y también lo están otros que solamente contienen memoria; así pues, pocas serían las innovaciones necesarias para producir una tarjeta que incorporase las funciones descritas sobre microcircuitos independientes. Podría resultar mucho más caro diseñar y fabricar una única pastilla que contuviera la totalidad de los microcircuitos necesarios.

Hay, sin embargo, dos razones imperiosas para alojar todos los dispositivos electrónicos en una sola pastilla. En primer lugar, las probabilidades de fallo resultan ser de casi el doble al utilizar dos pastillas en las situaciones de malos tratos físicos que sufre una tarjeta de crédito (que es doblada, retorcida, sometida a salpicaduras, vertidos de líquidos y temperaturas extremas) en lugar de una. Cada una de las dos pastillas podría fallar con pareja facilidad que una sola, y bastaría que una de las dos fracasara para dejar inutilizada la tarjeta. En segundo lugar, si la memoria se encuentra separada de la unidad central de procesamiento tiene que haber entre ambas una conexión eléctrica. Y sería posible para terceros supervisar los enlaces, violando así la seguridad de la memoria de la pastilla.

Lo peculiar del ambiente en que han de funcionar los circuitos microelectrónicos de la tarjeta inteligente no sólo impone estrictas restricciones al diseño global de la tarjeta, sino también sobre el diseño de la pastilla. Por ejemplo, un microcircuito inserto en una tarjeta de crédito tendrá que ser más delgado que la oblea ordinaria de silicio. Los microcircuitos diseñados para las tarjetas inteligentes tienen un espesor de 28 centésimas de milímetro en lugar de las 37 que son habituales. Además, la superficie de la pastilla debería ser tan diminuta como fuera posible: cuanto más pequeña, tanto menores los esfuerzos mecánicos que habrá de soportar cuando se doble la tarjeta de plástico. Por consiguiente, la selección de microcircuitos de tipos que puedan ser muy densamente empaquetados en la pastilla reviste particular importancia. También el costo de producción de la pastilla puede ser importante.

El microcircuito debería poder funcionar bien en ambientes electrónicamente "ruidosos". Por ejemplo, una tarjeta de cargo que vaya a utilizarse en un surtidor de gasolina tiene que poder funcionar incluso cerca de otros muchos dispositivos electrónicos, como los aparatos electromecánicos de automóviles cercanos y del propio surtidor. Importa también que los microcircuitos no consuman demasiada energía eléctrica. Sin embargo, en muchas aplicaciones no es necesario que la unidad central de procesamiento sea especialmente rápida, pues al manejar la tarjeta casi todas las operaciones a realizar comportan un diálogo con el usuario, y hasta los más lentos microprocesadores tienen tiempos de respuesta suficientes para ir al paso de la relativa lentitud de un humano al pulsar un teclado. (En ciertas aplicaciones, como las de alta seguridad, que comportarían amplios procesos de codificación y descodificación criptográfica, quizá sí convendría que se alcanzaran altas velocidades.)

Así pues, el ingeniero proyectista debe elegir circuitos de pequeño tamaño y que sea posible empaquetar densamente en una pastilla. Tiene que consumir la mínima energía eléctrica posible. Y tiene que ser barato de fabricar.

Quizás la decisión más importante consista en determinar qué tipos de transistores emplear. Los transistores son los ladrillos con que se construyen microprocesadores y memorias. Una pastilla ("chip") actual puede contener hasta 100.000 transistores. Estos cumplen en ellas papel de interruptores; cuando se aplica a uno de sus elementos una tensión adecuada, se abre o se cierra un canal eléctrico que conecta otros dos elementos. Cuando el canal está abierto, a través del dispositivo puede fluir una corriente.

Un transistor está formado por varias regiones adyacentes de un cristal de silicio deliberadamente impurificado ("dopado"), esto es, un cristal al que se han añadido átomos de impurezas para modificar sus propiedades eléctricas. Si los átomos de impureza añadidos tienen en su capa más externa (la llamada capa de valencia) más electrones que los átomos de silicio, habrá algunos electrones que no participarán en los enlaces que mantienen la estructura reticular cristalina, y que podrán moverse con una cierta libertad por el seno del silicio. Al aplicar al cristal un campo eléctrico, los electrones libres se desplazarán, creando una corriente eléctrica. Dado que los electrones que portan la corriente están cargados negativamente, cuando en el silicio se han introducido átomos que ceden electrones, se dice que está impurificado negativamente, y se le llama silicio *n*.

Recíprocamente, pueden aportarse átomos de impureza que tengan menos electrones de valencia que el silicio. Habrá entonces en el seno del cristal cierto número de "huecos", es decir, plazas electrónicas vacantes, que normalmente estarían ocupadas. En cierto sentido, los huecos actúan como las partículas positivamente cargadas: cuando se aplica al cristal un campo eléctrico, los huecos se desplazan de unos átomos a otros, produciendo una corriente eléctrica. Dado que esta corriente está formada por "partículas" positivamente cargadas, se dice que el silicio con exceso de huecos tiene una impurificación (o dopado) de tipo *p*.

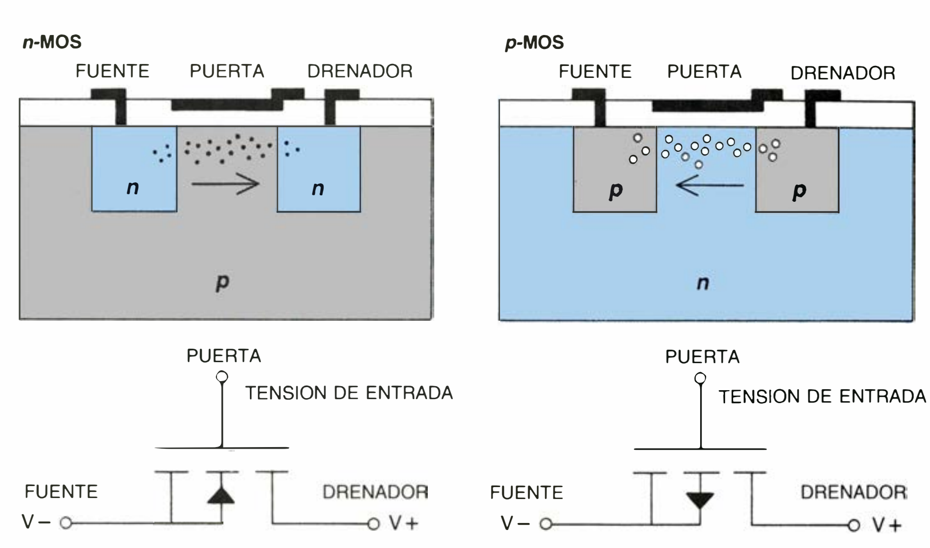
En el tipo de transistor conocido por bipolar *npn*, se tiene una región con una fuerte impurificación de tipo *n*, llamada colector, en el seno de una región con ligera impurificación del mismo tipo [*véase la figura 3*]. En esta misma región de ligera impurificación *n* se encuentra también una región con impurificación *p*, llamada base. Una tercera región, fuertemente impurificada, de tipo *n*, llamada emisor, yace en el seno de la región de tipo *p*, y se encuentra totalmente rodeada por ella.

En general, los electrones no pueden fluir desde el emisor al colector, ni siquiera aunque se aplique al colector

3. UN TRANSISTOR BIPOLAR *NPN* actúa como un interruptor. Al aplicar a uno de sus elementos, llamado base, una tensión positiva, puede fluir una corriente entre otros dos electrodos, llamados emisor y colector. El transistor se fabrica con silicio "dopado", que es silicio deliberadamente impurificado con átomos de otros tipos, a fin de modificar sus propiedades electrónicas. El silicio *n* dispone de electrones excedentarios, que pueden desplazarse formando una corriente eléctrica. En el silicio de tipo *p*, por el contrario, hay "huecos", puestos vacantes que pueden ocupar los electrones. Los huecos, que se comportan como si fueran partículas positivamente cargadas, se mueven también y forman corrientes eléctricas. En un transistor bipolar *npn* la corriente no puede, en principio, fluir desde el emisor hasta el colector, porque los electrones no pueden penetrar en la región de base, que es de tipo *p*. Sin embargo, al aplicar a la base y al colector tensiones positivas adecuadas, fluyen de la base al emisor algunos huecos, repelidos por la tensión positiva. A cambio, algunos electrones pasan desde el emisor hasta la base. La mayoría de estos electrones se difunden por la base y penetran en la región débilmente impurificada de tipo *n* que la rodea, y acaban por llegar al colector. En los esquemas de conexionado se utiliza para el transistor el símbolo que vemos abajo. Por convenio, la corriente suele presentarse siempre mediante flechas que apuntan en sentido contrario al del flujo electrónico.

una tensión positiva con respecto al emisor (una tensión que atrae electrones); no pueden atravesar la región de base, de impurificación *p*. Sin embargo, al aplicar a la base una tensión positiva, pasan algunos huecos desde la base hasta el emisor; a cambio, se inyectan en la base algunos electrones procedentes del emisor. Muchos de estos electrones atraviesan la base y penetran en la región (substrato) ligeramente impurificada de tipo *n*, en la cual yacen el colector, el emisor y la base. Desde allí pasan al colector. De este modo, la tensión positiva que le fue aplicada a la base actúa como señal de entrada, e induce el flujo de una corriente entre el emisor y el colector. Esta corriente es la señal de salida del transistor.

Los transistores bipolares son relativamente baratos de fabricar, y como conmutadores son rápidos (es decir, responden prestamente cuando se aplica a la base una tensión). Por otra parte, consumen bastante energía, pues cuando el transistor está en conducción (lo que sucede al aplicar a la base una tensión positiva) fluye corriente a su través, tanto desde la base como desde el emisor. Así pues, los transistores bipolares no son del todo adecuados para muchas de las aplicaciones que se darán a las tarjetas inteligentes.



4. TRANSISTORES MOS (metal-óxido-semiconductor): pueden construirse de dos modos. En un transistor *n*-MOS (*a la izquierda*) no puede fluir corriente a través de la región de silicio *p* que separa a la fuente del drenador, a menos que se aplique una tensión positiva a una puerta situada justamente encima de una delgada película de dióxido de silicio, que es aislante. Esta tensión positiva atrae una fina capa de electrones, llamada capa de inversión, que forma un canal conductor a través del cual puede fluir la corriente. Por el contrario, en un transistor *p*-MOS (*a la derecha*) para formar un canal debe aplicarse a la puerta una tensión negativa, que atrae una capa de inversión formada por huecos. Los huecos fluyen desde el drenador hacia la fuente.

Mucho más prometedores que los transistores bipolares son los construidos con tecnología MOS (metal-óxido-semiconductor). Hay dos tipos de transistores MOS: *n*-MOS y *p*-MOS. Un transistor *n*-MOS contiene dos islotes de silicio fuertemente impurificado, de tipo *n*, llamados fuente y drenador; ambos islotes están incrustrados en un sustrato de material de tipo *p* [*véase la figura 4*]. Una delgada capa de dióxido de silicio (que es aislante) cubre el cristal de silicio; un electrodo metálico, o de silicio, llamado puerta, es depositado después encima de la región de dióxido de silicio que se encuentra justamente sobre la porción de silicio *p* situada entre las dos regiones de impurificación *n*.

Lo mismo que en los transistores bipolares, la corriente no puede normalmente pasar a través del dispositivo; para pasar desde uno de los islotes de tipo *n* hasta el otro tendría que atravesar una región de tipo *p*. Sin embargo, al aplicar a la puerta una diferencia de potencial positiva, algunos electrones del sustrato de tipo *p* son atraídos hacia la región de silicio de tipo *p* situada justamente debajo de la capa de dióxido de silicio, entre la fuente y el drenador. Estos electrones forman una delgada capa; se la denomina capa de inversión porque está compuesta por portadores de carga de signo opuesto al de los huecos normalmente residentes en un semiconductor de tipo *p*. La capa de inversión se comporta como un canal de conducción, por donde fluyen otros electrones, desde la fuente hasta el drenador.

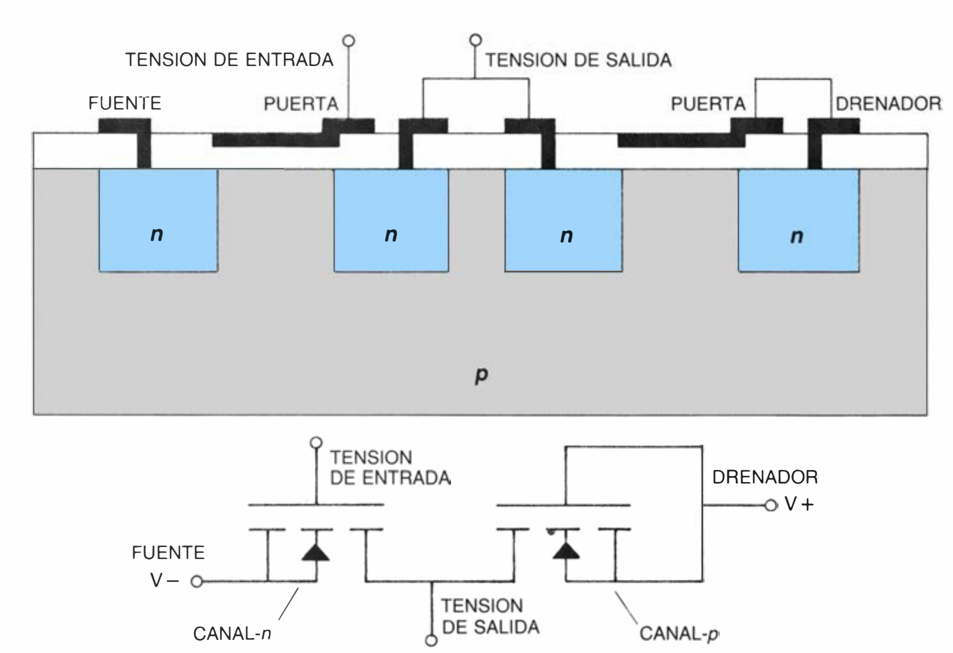
Los transistores *p*-MOS funcionan según el mismo principio que los *n*-MOS, pero la fuente y el drenador son de tipo *p* y yacen en un substrato de material de tipo *n*. En los transistores *p*-MOS, los portadores de corriente son los huecos que pasan del drenador a la fuente, y no los electrones; para atraer huecos hasta las proximidades de la superficie del transitor, y formar el canal entre el drenador y la fuente, se requiere una tensión negativa, a diferencia del caso anterior, en que era positiva.

En los circuitos MOS, lo mismo que en los transistores bipolares, la tensión aplicada a la puerta representa una señal de entrada. Sin embargo, a diferencia de los transistores bipolares, ahora no se utiliza como señal de salida la corriente que pasa entre la fuente y el drenador. En casi todos los circuitos de tecnología MOS, el drenador está eléctricamente conectado a una línea de tensión elevada, sea a través de una resistencia, sea mediante un transistor cuyo canal es siempre conductor [*véase la figura 5*]. La fuente está conectada, a su vez, a una tensión baja.

Cuando la tensión de entrada (que es la tensión aplicada a la puerta del transistor) es baja, el transistor no puede conducir corriente y el drenador no tiene conexión eléctrica directa hasta la línea de tensión baja. Sin embargo, el drenador sí está directamente conectado a una tensión alta. Por consiguiente, su tensión es alta. En cambio, cuando la señal de entrada es una tensión alta, el transistor puede conducir corriente; el drenador queda entonces conectado a la línea de baja tensión, y su tensión desciende. El potencial del drenador, que es alto cuando la tensión de entrada es baja, y bajo cuando la tensión de entrada es alta, constituye la señal de salida del dispositivo MOS.

Estos circuitos MOS son razonablemente rápidos, baratos de fabricar y fiables en su funcionamiento. Se empaquetan densamente en una pastilla ("chip"). Sin embargo, son relativamente sensibles al ruido (señales eléctricas generadas por dispositivos vecinos), y aunque su funcionamiento exige menos energía que los transistores bipolares, todavía absorben mucha. Por consiguiente, pudieran resultar inaceptables en algunas aplicaciones de las tarjetas inteligentes.

Una variante de esta clase de tecnología MOS, la llamada MOS complementaria, o CMOS, requiere un consumo energético inferior en aproximadamente un orden de magnitud, y su sensibilidad al ruido viene a ser la mitad. Posiblemente resulte ser la tecnología adecuada para las tarjetas inteligentes.



5. UN DISPOSITIVO CONMUTADOR MOS de canal *n* consta de dos transistores *n*-MOS interconectados. En uno de los transistores se conectan entre sí la puerta y el drenador (*a la derecha*); estos elementos se conectan, a su vez, a una línea de tensión alta, por lo cual en ese transistor puede fluir corriente desde la fuente al drenador. La fuente del otro transistor (*a la izquierda de todo*) está conectada a una línea de tensión baja. Su drenador está directamente conectado a la fuente del primer transistor; la tensión eléctrica de este nodo (*centro*) representa la tensión de salida del dispositivo. Cuando la tensión de entrada (*izquierda*) es baja, el nodo de salida está eléctricamente aislado de la fuente del transistor de la izquierda, pues no existe capa de electrones que pueda conducir una corriente. Sin embargo, el nodo de salida está conectado, a través del transistor permanentemente conductor, a la línea de tensión alta conectada al drenador del dispositivo. Por tanto, el nodo de salida toma una tensión elevada. Cuando la tensión de entrada es alta, el nodo de salida deja de estar aislado de la línea de tensión baja conectada al dispositivo y adquiere una tensión baja.

Un circuito CMOS está formado por un único cristal que contiene a un tiempo un transistor *n*-MOS y otro *p*-MOS. Las puertas de ambos transistores están conectadas entre sí [*véase la figura 6*]; la tensión aplicada a ambas actúa de señal de entrada. Cualquiera que sea la señal de entrada, las respuestas de ambos transistores serán complementarias, pues una señal que active al transistor *p*-MOS desactivará al *n*-MOS, y viceversa.

El drenador del transistor *n*-MOS se halla conectado directamente con la fuente del transistor *p*-MOS, por cuya razón estos elementos están siempre a la misma tensión. Esta tensión común es la señal de salida. La fuente del transistor *n*-MOS queda conectada a una línea de tensión baja, y el drenador del transistor *p*-MOS a una línea de tensión elevada.

Cuando la tensión de entrada es positiva, el canal del transistor *n*-MOS se hace conductor, pero no el del *p*-MOS. El drenador del transistor *n*-MOS y la fuente del transistor *p*-MOS se encuentran, pues, conectados a la línea de tensión baja y aislados de la línea de tensión alta. Por consiguiente, la tensión de salida tiene un valor bajo. Recíprocamente, cuando la tensión de entrada es negativa, la corriente puede fluir a través del transistor *p*-MOS, pero no del transistor *n*-MOS. El drenador del transistor *n*-MOS y la fuente del *p*-MOS se encuentran entonces conectados a la línea de tensión elevada, y aislados de la línea de tensión baja. Por tanto, la tensión de salida toma un valor alto.

Seguramente, la característica más importante de los dispositivos CMOS sea que no pase corriente entre la línea de tensión alta conectada al drenador del transistor *p*-MOS y la línea de tensión baja, conectada a la fuente del transistor *n*-MOS, salvo durante los breves períodos en que la señal de la entrada es conmutada entre una tensión alta y una baja. Así pues, los dispositivos CMOS absorben corrientes mucho menores que los transistores bipolares o que un par de transistores que sean ambos *n*-MOS o ambos *p*-MOS.

Igualmente, los dispositivos CMOS son mucho menos sensibles al ruido eléctrico. Para conmutar un dispositivo *n*-MOS normal desde, pongamos por caso, una tensión de salida baja a una tensión de salida alta, lo único que hace falta es conmutar la tensión de entrada desde cierto valor positivo, en cuyo caso el canal conductivo entre fuente y drenador está abierto, a tensión cero. (Cuando la tensión de puerta se hace igual a cero dejan de ser atraídos electrones hasta la superficie del dispositivo, y no existe canal conductor.) Para conmutar un dispositivo CMOS es preciso conmutar la tensión de entrada desde un valor positivo, que permite la conducción del transistor *p*-MOS, hasta un valor negativo, que hace conducir al transistor *n*-MOS. La diferencia entre la nueva tensión de entrada y la antigua es, en un dispositivo CMOS, de valor doble que para circuitos *n*-MOS y, por tanto, las señales espúreas tendrían que ser de intensidad doble para inducir un error en el dispositivo CMOS.

Los circuitos CMOS son en la actualidad más caros de fabricar que los de tecnología *n*-MOS, y no admiten tan gran densidad de empaquetamiento en una pastilla como éstos. Sin embargo, la razón de que así suceda débese, en parte, a la relativa novedad de la tecnología CMOS. Dentro de pocos años será posible fabricar dispositivos CMOS tan pequeños y baratos como los *n*-MOS. Existen ya algunos prototipos de esta clase de dispositivos, llamados CMOS de alto rendimiento, o HCMOS (de "high-performace CMOS"). La tecnología de los dispositivos HCMOS es esencialmente idéntica a la de los CMOS, pero el equipo de fabricación es más preciso, lo que permite incluir más componentes en cada unidad de superficie. La separación entre la fuente y el drenador de un dispositivo CMOS es, típicamente, de unos tres micrometros; la de un dispositivo HCMOS, de menos de un micrometro.

Hay una última cuestión tecnológica que concierne a la memoria de la tarjeta inteligente. Cada tarjeta inteligente habrá de tener al menos tres tipos generales de memoria. Una de éstas, llamada "memoria de sólo lectura" (ROM), es creada en fábrica y no puede ser alterada. En ella está contenido, por ejemplo, el sistema operativo de la tarjeta, esto es, el conjunto de instrucciones que ha de seguir la unidad central de procesamiento (la parte de la pastilla encargada de efectuar las operaciones lógicas) cuando se le dan determinadas instrucciones. Otra memoria, una memoria volátil de acceso aleatorio, o RAM, es una memoria muy rápida que el fabricante deja en blanco. Cualquier porción de la RAM puede ser alterada por la unidad central de procesamiento de la tarjeta, pero la información de la RAM queda borrada en cuanto la tarjeta es desconectada de

una fuente de alimentación. La memoria volátil se utiliza como si fuera un bloc de notas: sirve para almacenar los resultados intermedios de los cálculos realizados por la unidad central de procesamiento.

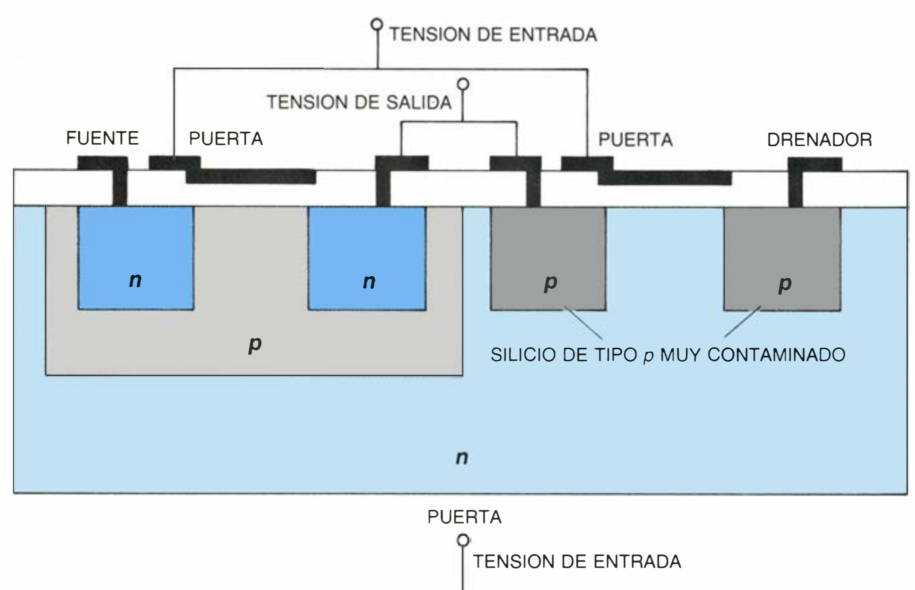
La mayor parte del tercer tipo de memoria, la memoria no volátil, programable, de lectura únicamente, es dejada en blanco por el fabricante y puede modificarse por la unidad central de procesamiento. La información introducida en esta memoria permanece en ella incluso después de retirar la tarjeta de la fuente de alimentación. Esta es la memoria que aloja las zonas de trabajo, la zona abierta y la zona secreta, que ayudan a garantizar la seguridad de la información almacenada en una tarjeta inteligente.

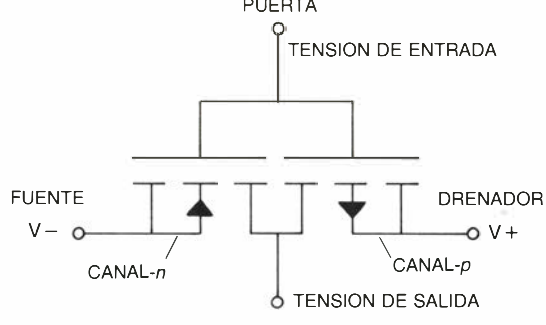
Hasta recientemente, la única tecnología que permitía construir este tipo de memorias era la llamada memoria EPROM (en inglés, "erasable, programmable, read-only memory", o sea, memoria únicamente de lectura, borrable y programable). El nombre puede confundir un poco, pues la información introducida en una memoria EPROM solamente puede borrarse mediante exposición a luz ultravioleta, bajo condiciones controladas. El titular tendría que devolver la tarjeta al proveedor, a fin de que borrase la memoria y la tarjeta pudiera volver a ser utilizada. Una tarjeta que se funde en la técnica EPROM sirve durante un período limitado, al cabo del cual ya no habrá disponible más espacio de memoria para almacenar informaciones tales como la fecha y el importe de lo adquirido. Lo ideal sería que el titular de la tarjeta pudiera borrar de la memoria las transacciones del año anterior para hacer sitio para las del siguiente.

Un tipo nuevo de dispositivo de memoria, conocido por memoria de sólo lectura, programable y electrónicamente borrable (EEPROM, o "electronically erasable r.o.m.") puede que pronto haga posible tal realidad. Una célula de EEPROM se construye en un único cristal de dióxido de silicio aislante, que yace sobre un substrato de silicio *p*. Debajo mismo de la superficie superior del cristal hay un islote de polisilicio (que es buen conductor), llamado puerta de control [*véase la figura* 7]. La puerta de control está eléctricamente conectada a un electrodo de control situado por encima del cristal de dióxido de silicio. Debajo de la puerta de control hay otra isla de polisilicio, llamada puerta flotante, que está eléctricamente aislada de todos los demás componentes de la célula EEPROM. La puerta flotante está separada del substrato por una delgada capa de dióxido de silicio. En el seno del substrato, sobre cada lado de la región que yace directamente debajo de la puerta flotante, se encuentran la fuente y el drenador, dos islotes de silicio impurificado de tipo *n*.

La célula EEPROM se funda en una propiedad de carácter cuántico de los electrones, llamada "efecto túnel". Según las reglas de la mecánica cuántica, cuando se encuentra presente un electrón en un material conductor que está separado de otro material conductor por una delgada capa de material aislante, existe una pequeña probabilidad de que el electrón atraviese por "efecto túnel" la delgada lámina y quede atrapado en el segundo conductor. En el caso de una EEPROM, los electrones se desplazan, de acuerdo con este efecto, entre el silicio de impurificación *p* y la puerta flotante de polisilicio. Es la presencia o ausencia de electrones supernumerarios en el seno de la puerta flotante la que denota si una célula EEPROM está almacenando el dígito binario 1 o 0.

Cuando no se le aplica tensión alguna a la fuente, el drenador y la puerta, no se dispone de electrones que puedan trasladarse por efecto túnel. Para producirlos hay que aplicar a la fuente un pequeño voltaje positivo y un gran voltaje positivo a la puerta de control. La fuente, el drenador y la puerta de control se comportan como sus homónimos de un transistor *n*-MOS: una capa de electrones es atraída a la superficie del substrato impurificado de tipo *p*, y a través de esta capa fluye una corriente, desde la fuente hasta el drenador. Algunos de los electrones que





6. DISPOSITIVOS CONMUTADORES CMOS (o sea, MOS complementarios): consumen mucha menos energía que los conmutadores *n*-MOS o *p*-MOS estándar; podemos ver en ellos la tecnología más apropiada para las tarjetas inteligentes. En los dispositivos CMOS se conectan entre sí las puertas de un transistor *p*-MOS y de otro *n*-MOS; la tensión simultáneamente aplicada a ambas es la señal de entrada. Las acciones de los dos transistores serán, consiguientemente, complementarias: siempre que una tensión active al transistor *n*-MOS, la misma tensión desactivará al *p*-MOS, y viceversa. Además, el drenador del transistor *n*-MOS está conectado a la fuente del transistor *p*-MOS; la tensión de este nodo es la señal de salida del dispositivo. La fuente del transistor *n*-MOS está conectada a una línea de tensión baja, y el drenador del *p*-MOS a una línea de tensión alta. Cuando la tensión de entrada sea positiva podrá fluir corriente a través del transistor *n*-MOS, pero no a través del transistor *p*-MOS; el nodo de salida, que de este modo queda conectado eléctricamente a una línea de tensión baja, y que se encuentra aislado de la línea de tensión alta, tendrá una tensión baja. Recíprocamente, cuando la tensión de entrada sea negativa, el nodo de salida estará conectado a una línea de tensión alta por intermedio del transistor *p*-MOS, y aislado de una línea de tensión baja; por consiguiente queda a tensión alta. Los dispositivos CMOS consumen poquísima potencia, porque a través del dispositivo no fluye corriente alguna, excepto en los breves períodos en que la tensión de entrada pasa de alta a baja, o viceversa.

componen esta corriente se cuelan, por efecto túnel, en la puerta flotante y quedan atrapados allí. Al suprimir las tensiones del drenador y de la puerta de control, los electrones permanecen en la puerta flotante y la EEPROM almacena un bit denotativo de un 1 lógico.

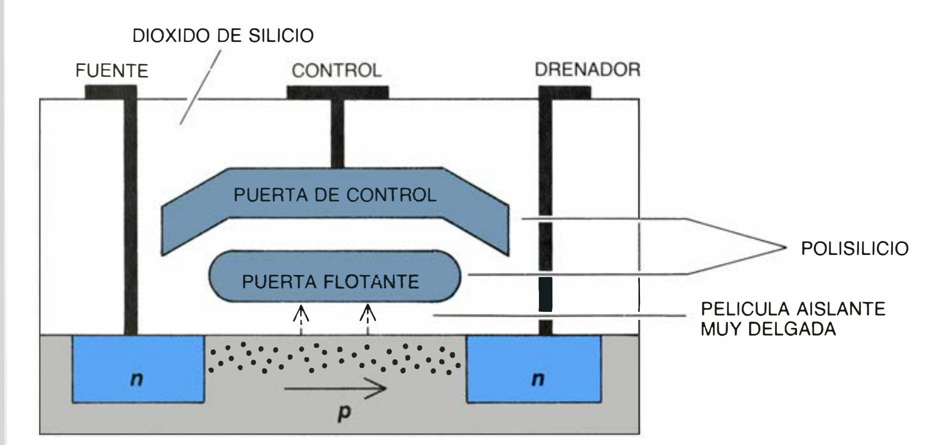
Para eliminar los electrones atrapados en la puerta flotante, se aplica al drenador una pequeña tensión positiva, al tiempo que se le aplica a la puerta de control un elevado voltaje negativo. Los electrones, repelidos entonces por la puerta de control, vuelven a pasar, por efecto túnel, hasta el substrato, donde forman una corriente entre la célula y el drenador.

No es posible programar y reprogramar indefinidamente una EEPROM. Cada vez que los electrones se infiltran por efecto túnel entre el substrato y la puerta flotante, quedan atrapados unos cuantos en el interior de pequeños defectos de la finísima capa aislante de dióxido de silicio. Al cabo de unos 10.000 ciclos (es decir, después de que la EEPROM se haya reprogramado de 0 a 1 y viceversa 10.000 veces) quedan atrapados suficientes electrones como para crear un canal de fugas a través de los cuales la corriente fluye directamente desde la puerta flotante hasta el substrato.

Sin embargo, una tarjeta inteligente que estuviera equipada con memoria EEPROM sería lo suficientemente duradera como para ser utilizada durante varios años, y tendría una enorme flexibilidad de funcionamiento.

La tarjeta inteligente puede provocar cambios fundamentales en el funcionamiento económico de la sociedad. Por ejemplo, utilizadas con el papel de llaves, las tarjetas inteligentes pueden proporcionar el grado de seguridad necesario para hacer que las redes de ordenadores sean verdaderamente viables. Para que pueda funcionar un sistema totalmente electrónico de compensaciones bancarias y de transferencias de fondos resulta imprescindible garantizar que no tengan acceso al sistema usuarios no autorizados. Las tarjetas inteligentes pueden crear esta auténtica unión entre informática y telecomunicaciones.

Las tarjetas inteligentes cambiarán también la forma en que se realizan las transacciones comerciales sencillas. Ha transcurrido ya algún tiempo desde que el sistema del trueque, es decir, del intercambio directo de un bien por otro, fuera reemplazado por el sistema de cambiar bienes por unidades normalizadas de riqueza (ya sean conchas o piezas de oro). Al cabo, las propias unidades de riqueza fueron reemplazadas por recibos, es decir, certificados que representaban una cierta cantidad de oro o plata. Actualmente, aquellos certificados han sido reemplazados por los billetes y los cheques bancarios. En un futuro, el papel moneda puede quedar reemplazado por "unidades de poder adquisitivo", almacenadas electrónicamente en tarjetas inteligentes.



7. CELULAS DE MEMORIA que integran una memoria de sólo lectura, programable y borrable eléctricamente (EEPROM) retienen la información almacenada en ellas después incluso de haber sido desconectadas de la fuente de alimentación. La presencia o ausencia de electrones en una "puerta flotante", eléctricamente aislada (*centro de la figura*), expresa si la célula está almacenando un dígito binario 1 o un 0. Para llevar electrones hasta la puerta flotante, y dejarlos atrapados allí, se aplica al drenador una pequeña tensión positiva, y una gran tensión positiva a la puerta de control. La capa de electrones así atraída hasta la zona situada bajo la puerta flotante permite el paso de una pequeña corriente entre la fuente y el drenador. Algunos de estos electrones, atraídos por la tensión positiva aplicada a la puerta de control, se "cuelan" por efecto túnel a través de una delgada película de dióxido de silicio, introduciéndose en la puerta flotante, y quedando atrapados allí. Para sacarlos se aplica una fuerte tensión negativa a la puerta de control, al propio tiempo que una pequeña tensión positiva al drenador. Los electrones vuelven entonces a pasar, por efecto túnel, desde la puerta flotante hasta el sustrato, que es de silicio *p*, donde forman una pequeña corriente que va hasta el drenador. De este modo, las células EEPROM pueden programarse y reprogramarse muchas veces.

Comportamiento de los líquidos en ingravidez

El estudio de los líquidos en condiciones de baja gravedad constituye un paso previo para el desarrollo de procesos de manufactura que permiten un uso comercial del espacio

José Meseguer Ruiz y Angel Sanz Andrés

A finales de 1983, entre los días 28 de noviembre y 8 de diciembre, tuvo lugar la primera misión del laboratorio espacial europeo *Spacelab 1.* Durante ese vuelo se realizaron a bordo de la nave un elevado número de experimentos; en particular, uno relacionado con el comportamiento de líquidos en condiciones de baja gravedad. En esa situación, la respuesta del líquido está gobernada, fundamentalmente, por las fuerzas de tensión superficial.

Desde 1974, financiado por la Comisión Nacional española de Investigación del Espacio, se ha venido desarrollando en el Laboratorio de Aerodinámica de la Escuela Técnica Superior de Ingenieros Aeronáuticos de la Universidad Politécnica de Madrid, bajo la dirección de I. Da Riva, un programa de investigación básica sobre columnas líquidas en condiciones de ingravidez, cuya finalidad es el análisis de la respuesta frente a perturbaciones mecánicas de una columna líquida en condiciones de baja gravedad.

El interés de tales estudios, aparte de las aplicaciones industriales que se detallan posteriormente, reside en la esperanza de una utilización comercial del espacio, por lo que la investigación de procesos de manufactura en condiciones de baja gravedad es, hoy por hoy, una de las más atrayentes empresas científicas, tendente a aumentar nuestro conocimiento sobre ciertos fenómenos que en la vida cotidiana permanecen enmascarados por la gravedad.

En gravedad reducida casi desaparece la convección inducida por la flotabilidad, apenas existe sedimentación y es posible mover volúmenes grandes de líquido mediante las fuerzas de tensión superficial. Por supuesto, no todo son ventajas en el manejo de líquidos en ingravidez, y entre los inconvenientes podríamos contar el que la transmisión de calor es menos eficaz, la dificultad de desgasificar masas fundidas y la aparición de corrientes de convección generadas por gradientes de tensión superficial. Pero, ¿en qué consiste la tensión superficial?

El hecho de que las pequeñas gotas de un líquido en el aire, o las burbujas de aire en el agua, adopten una forma esférica se justifica admitiendo la hipótesis de que la superficie de separación entre dos fluidos en equilibrio, la entrefase, es la base de una forma especial de energía proporcional al área de dicha entrefase. Cuando se establecen las relaciones termodinámicas que determinan la evolución de un sistema fluido, se suele hablar en términos de energía y trabajo por unidad de masa, suponiendo implícitamente que la cantidad total de energía y trabajo son, en todos los casos, proporcionales al volumen del fluido considerado. Esto no es cierto, sin embargo, en aquellos casos en los que se considera una masa fluida con una relación grande entre superfice y volumen, o bien en los que las fuerzas másicas no existen o son pequeñas; en estos casos hay que tener muy en cuenta la contribución a la energía total del término dependiente de la entrefase.

Así, se admite que la entrefase de un sistema fluido de volumen constante en equilibrio contribuye a la energía libre de Helmholtz del sistema con una cantidad proporcional al área de la entrefase, siendo la constante de proporcionalidad, σ, una función de estado del sistema.

Si suponemos ahora que el sistema experimenta una evolución reversible a temperatura constante, de tal modo que tanto la densidad como el volumen de las fases que lo integran permanecen constantes, el trabajo total realizado será el producto de σ por la variación del área de la entrefase. Esta expresión del trabajo que se debería hacer sobre el sistema para cambiar sólo el área de la entrefase es la misma que se obtendría suponiendo que la entrefase es una membrana sometida a un estado de tensión uniforme. En consecuencia, la función de estado σ puede ser interpretada bajo dos puntos de vista: como la energía de la entrefase, por unidad de área, o como una "tensión superficial", en el sentido de que a lo largo de cualquier línea trazada sobre la entrefase existe una fuerza por unidad de longitud, de magnitud σ, cuya dirección es normal a la línea y tangente a la entrefase.

El fenómeno de la tensión superficial tiene un origen molecular cuya explicación requiere el concurso de las fuerzas de cohesión intermoleculares. En efecto, la energía libre media asociada a una molécula en un cierto medio fluido es independiente de la posición de la molécula en el interior del fluido; sin embargo, a distancias de la entrefase del orden del radio de acción de las fuerzas intermoleculares, cifrado en 10^{-9} metros para moléculas simples, el valor de la energía libre media resulta afectado por la proximidad de la superficie. Cuando de los dos fluidos que configuran una entrefase uno es un líquido y el otro un gas, las moléculas del líquido próximas a la entrefase están sometidas a una acción atractiva ejercida por las moléculas vecinas del líquido, acción que no es contrarrestada por las moléculas del gas, mucho más dispersas; el resultado es una tendencia de las moléculas de la entrefase a moverse hacia el interior del líquido, es decir, a contraer la superficie minimizando el área de la entrefase. Cuando ello ocurre, la tensión superficial es positiva.

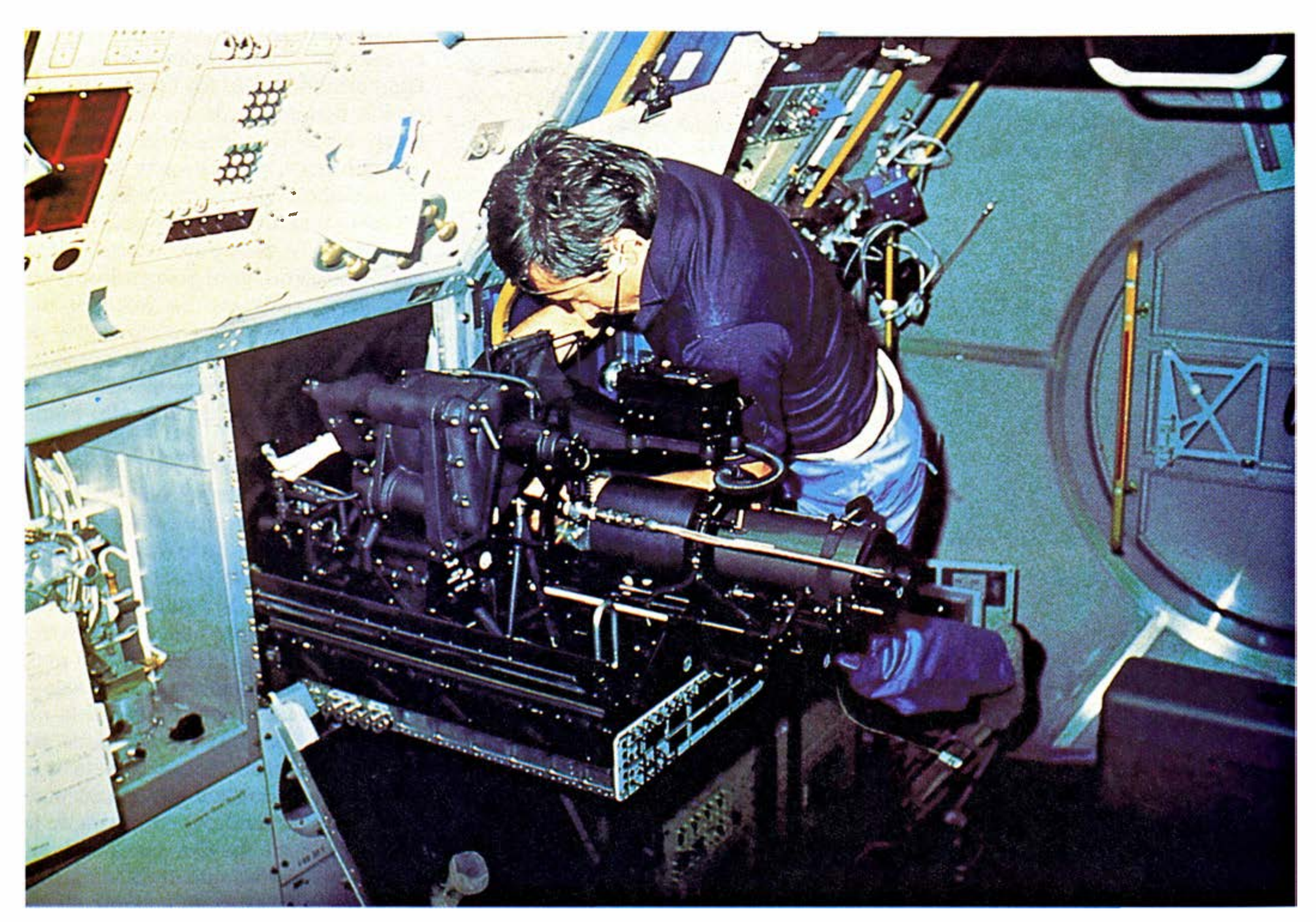
Cuando la entrefase separa dos líquidos, o un líquido y un sólido, el signo de la tensión superficial no es predecible mediante argumentaciones como la anterior y, de hecho, σ puede ser tanto positiva como negativa. Si para un cierto par de líquidos el valor de σ es negativo, el conjunto evolucionará de modo que el área de la entrefase aumente, hasta donde pueda, llegándose rápidamente a un completo mezclado de los dos líquidos; este es el caso para el par alcohol-agua. Así pues, para que dos líquidos sean inmiscibles entre sí, es preciso que la tensión superficial correspondiente sea positiva.

Normalmente, el valor de la tensión superficial disminuye con la temperatura y con la adición de contaminantes. Refiriéndonos al agua, ésta es la explicación de por qué el lavado en caliente es más efectivo que en frío, y también explica la acción de los detergentes y jabones, ya que estos agentes, al reducir el valor de la tensión superficial, facilitan el mojado de las superficies sólidas en contacto con el agua.

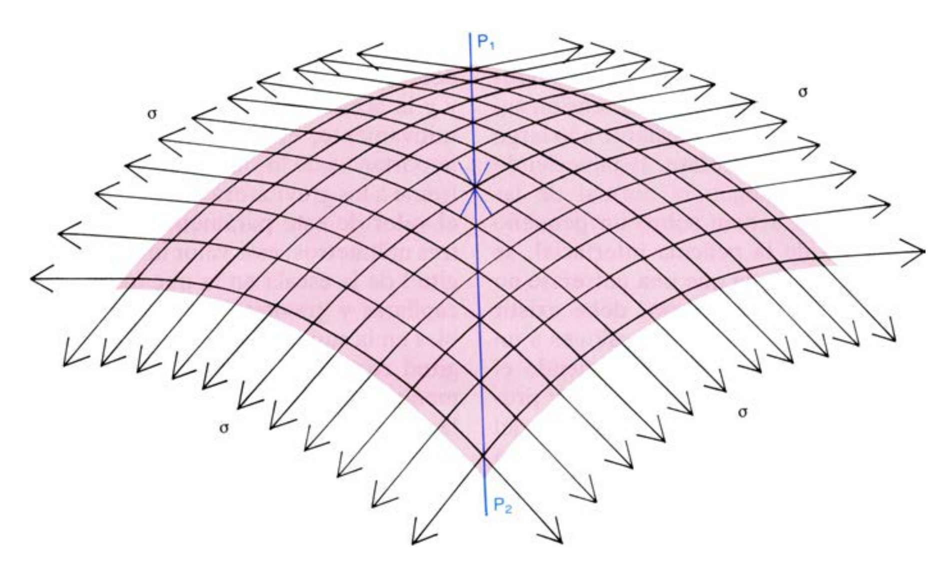
Volviendo a un punto de vista macroscópico, supongamos dos líquidos en equilibrio; el valor de la tensión superficial será, pues, constante a lo largo y ancho de la entrefase. Estableciendo el equilibrio según la normal de las fuerzas que actúan sobre un pequeño elemento de la película interfacial, se deduce que, para que una entrefase no plana esté en equilibrio, debe existir una diferencia entre las presiones a un lado y otro de la entrefase, siendo el salto de presiones proporcional al producto de σ por la curvatura media del elemento de superficie. Esta condición se modificará en el caso de una entrefase en movimiento, ya que deberán considerarse los esfuerzos viscosos a la hora de evaluar el salto de presiones.

Supongamos una entrefase aire-agua en reposo en la superficie terrestre; la condición de equilibrio en cualquier punto de la entrefase expresa la igualdad de las presiones hidrostática y capilar, y la comparación entre éstas permite establecer que el único parámetro importante en el problema, que tiene dimensiones de longitud, es $\sqrt{\sigma/\rho g}$, donde ρ designa la densidad del líquido y g la aceleración de la gravedad. (Intuitivamente, el parámetro expresa la importancia de los fenómenos capilares frente a los gravitatorios.) Para el agua, el valor de este parámetro es de unos tres milímetros; este valor indica la longitud de la escala en la que los efectos capilares y gravitatorios son comparables en la superficie terrestre. Si la longitud característica de la entrefase es menor, su forma estará gobernada por la tensión superficial, pero si es mayor, salvo efectos locales, será la gravedad la que determine la forma de la entrefase.

Normalmente, la entrefase entre dos fluidos termina en la línea a lo largo de la cual entran en contacto tres medios diferentes (piénsese, por ejemplo, en la línea de separación café-taza-aire). Esta línea triple está sujeta a las tensiones de las tres entrefases y, por tanto, para que pueda estar en equilibrio la resultante en cada punto de las diversas tensiones interfaciales debe ser nula. En el caso de que el valor, en



1. COMPORTAMIENTO FRENTE A PERTURBACIONES MECANICAS de una masa líquida mantenida, en ausencia de gravedad, entre dos discos sólidos por la sola acción de las fuerzas de tensión superficial (columnas líquidas en ingravidez). Objeto de estudio de un programa de investigación básica en desarrollo en el Laboratorio de Aerodinámica de la Escuela Técnica Superior de Ingenieros Aeronáuticos desde 1974. Un hito importante en este programa fue la realización de un experimento con puentes líquidos a bordo del Laboratorio Espacial Europeo Spacelab, durante la primera misión de esta nave a finales de 1983. En la fotografía se observa al astronauta alemán Ulf Merbold en el interior del Spacelab, manipulando el módulo de física de fluidos, un equipo diseñado especialmente para la experimentación con líquidos en ingravidez, en el que se realizó el experimento español junto con otros cinco de la misma área.

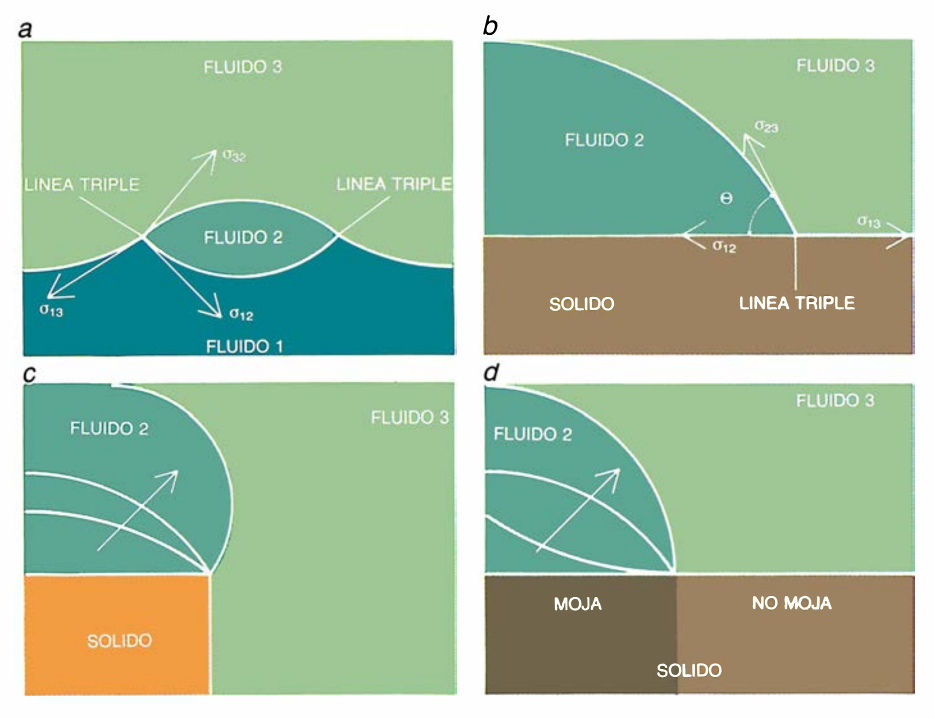


2. SUPERFICIE DE SEPARACION entre dos fluidos inmiscibles. Esa entrefase puede considerarse cual una membrana elástica sometida a una tensión σ cuyo valor, en un buen número de casos, es constante a lo largo de la entrefase. Sobre cualquier línea contenida en esa superficie actúa una fuerza por unidad de longitud de magnitud σ, cuya dirección es perpendicular al elemento de línea y tangente a la entrefase. Atendiendo a un elemento de superficie, la resultante de las fuerzas de tensión superficial según la normal a este área elemental es compensada por la diferencia entre las presiones a uno y otro lado de la entrefase.

módulo, de una de las tres tensiones superficiales sea superior a la suma de los módulos de los valores de las otras dos, no podrán alcanzarse las condiciones de equilibrio; esto es lo que ocurre cuando se vierte una gota de un aceite mineral sobre la superficie del agua: el valor de la tensión interfacial aire-agua es mucho más elevado que los correspondientes a los pares aire-aceite o aceite-agua; en consecuencia, la gota de aceite se extiende indefinidamente sobre la superficie del agua hasta cubrir toda la superficie libre o hasta que el espesor de la capa de aceite alcanza dimensiones moleculares. Si uno de los tres medios es un sólido, cuya superficie supondremos plana, la línea triple sólo puede moverse desplazándose a lo largo de esta superficie, y la condición de equilibrio dependerá del valor del llamado ángulo de contacto, θ. Cuando el sólido presenta una arista aguda, en esa arista cualquier ángulo de contacto es posible dentro de ciertas limitaciones: una arista es, pues, una barrera capaz de detener el movimiento de avance de la línea triple.

Este breve repaso al concepto de tensión superficial y sobre todo el haber establecido, siquiera en un caso particular, la escala en la que este fenómeno es importante en la superficie terrestre, puede servir para centrar nuestra atención sobre la multitud de procesos gobernados principalmente por la tensión superficial. Una simple taza de café constituye un pequeño laboratorio: el menisco que el café forma en las superficies de la taza, la ascensión capilar del café en el terrón de azúcar, la pequeña gota adherida a la cucharilla, etcétera; en todos estos fenómenos la tensión superficial desempeña un papel relevante. Mirando un poco más lejos, existe un gran número de técnicas industriales para las que es necesario un conocimiento preciso del comportamiento de los líquidos bajo la acción dominante de las fuerzas capilares, desde las técnicas de impresión por chorros capilares hasta los trenes de soldadura blanda, pasando por los procesos de fabricación de películas fotográficas; en particular, ciertas técnicas de punta basan su desarrollo en este fenómeno; tales son los métodos denominados EFG, de Czochralski y de la zona líquida, ampliamente utilizados en la industria de los semiconductores y de los materiales de muy alto punto de fusión.

En el método de la zona líquida se parte de una varilla del material a purificar sujeta en sus extremos, realizándose el proceso en vacío o en atmósfera inerte. Una fuente térmica, desplazable a lo largo de la varilla, la calienta localmente cerca de uno de sus extremos, al tiempo que la varilla gira para uniformizar las temperaturas; el calentamiento local prosigue hasta que la varilla funde. La zona fundida debe ser lo suficientemente pequeña para que las fuerzas de tensión superficial puedan retener al fundido, en contra de la gravedad, entre las dos partes sólidas en que ha quedado dividida la varilla; si ahora se desplaza lentamente la fuente térmica a lo largo de la misma, los frentes de fusión y de solidificación la recorrerán en toda su longitud y, en consecuencia, al finalizar el proceso toda la varilla se habrá fundido y soli-



3. EN UNA AMPLIA VARIEDAD DE SITUACIONES, la entrefase entre dos fluidos acaba donde éstos entran en contacto con un tercer medio, apareciendo una línea en la que confluyen tres entrefases diferentes (*a*). Para que esta línea triple pueda estar en equilibrio deberá ser nula la resultante de las tres fuerzas de tensión superficial que actúan en cada punto de la misma. Si uno de los tres medios es un sólido, la única posibilidad de desplazamiento de la línea triple se desarrollará a lo largo de la superficie del mismo (*b*). Ahora bien, cuando el sólido presenta una arista, son posibles distintos valores del ángulo de contacto, comportándose la arista como si fuera una barrera al avance de la línea triple (*c*). Una conducta análoga se observa cuando, aun siendo la superficie del sólido continuada, cambian las propiedades de mojado (*d*).

dificado de nuevo, aunque progresivamente. Dado que las impurezas tienden a permanecer en el fundido, gran parte de éstas habrán sido arrastradas hacia uno de los extremos de la barra y el resto del material tratado es ahora más puro que en la barra inicial.

El método presenta además otra ventaja; cuando se solidifica un fundido en crisol, existe un gran número de núcleos de solidificación a partir de los cuales se van formando los cristales sólidos por acreción de átomos desde el fundido; el resultado es un sólido de aspecto granular en el que cada grano presenta una estructura cristalina más o menos perfecta. Con la zona líquida se puede controlar el avance del frente de solidificación y obtener, por tanto, monocristales de gran tamaño.

El proceso descrito presenta, sin embargo, ciertas dificultades relacionadas con el tamaño de las muestras a tratar. En efecto, la presión hidrostática, que crece linealmente desde el borde superior del puente líquido, limita la longitud máxima de la zona fundida que apenas puede llegar a ser de unos pocos milímetros; en consecuencia, si el diámetro de la varilla es excesivo, puede ocurrir que la región superficial fundida sobrepase esta longitud crítica antes de que se haya fundido el núcleo de la barra. Así pues, para poder controlar con eficacia el proceso, el diámetro de la varilla debe ser del mismo orden de magnitud que la longitud de la zona fundida.

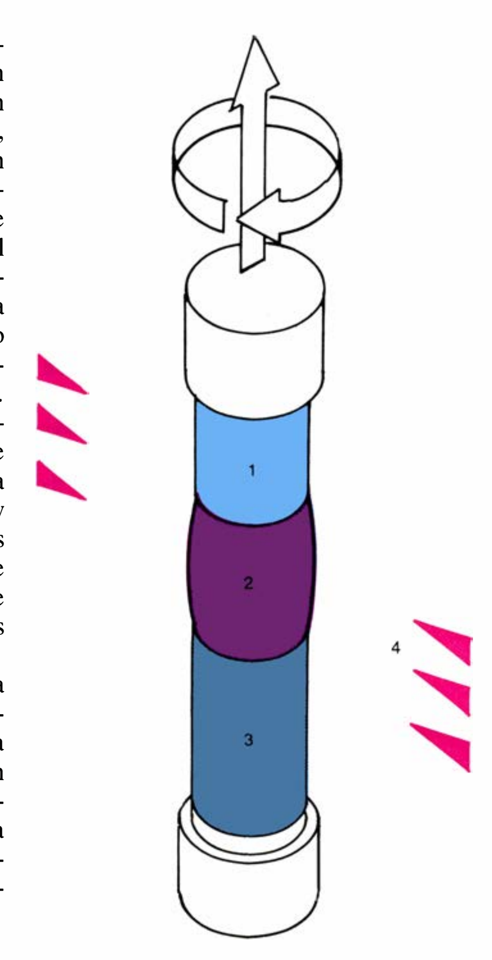
Esta limitación hizo que los científicos miraran hacia el espacio, sobre todo a raíz del lanzamiento del laboratorio espacial norteamericano Skylab, con la idea de que las dificultades desaparecerían en gran medida si se pudiera trabajar con zonas fundidas a bordo de una plataforma espacial en condiciones de gravedad reducida. En el espacio, la anulación de las fuerzas gravitatorias permitiría trabajar con muestras de gran tamaño, al desaparecer la severa restricción que limita la longitud máxima de la zona fundida en la superficie terrestre. (Esta afirmación debe aceptarse, sin embargo, con ciertas reservas, ya que la longitud máxima de una zona flotante está limitada por criterios de estabilidad.) En el espacio, las posibilidades de manejo y control son mayores y, además, aparte del interés industrial que ofrece la técnica de la zona flotante, existe también el interés científico de encontrar respuestas satisfactorias que puedan explicar el comportamiento de los líquidos en ingravidez, al menos para ciertas configuraciones particulares.

El estudio de la zona líquida en la configuración real antes descrita es tremendamente complicado. Intervienen frentes de fusión y de solidificación cuya geometría varía con el tiempo, fundidos cuyas propiedades dependen de la temperatura, gradientes espaciales de las propiedades, etc., por lo que resulta inabordable de un modo global con el estado de conocimientos actuales. Ante esta situación, la única salida es modelar un problema simplificado que incluya, al menos, las características más sobresalientes del caso real. Nuestro modelo se centra, fundamentalmente, en los aspectos mecánicos de la zona flotante, y consiste en una masa líquida, de propiedades uniformes y constantes, mantenida por las fuerzas de tensión superficial en ausencia de gravedad entre dos discos sólidos, de igual diámetro, coaxiales y paralelos entre sí.

El estudio de esa configuración, una columna líquida entre apoyos circulares, es el objeto de nuestro programa de investigación, en el que se pueden distinguir tres grandes áreas: investigación teórica, experimentación a bordo de naves espaciales y experimentación en tierra en microgravedad simulada.

Cronológicamente, los estudios teóricos se iniciaron con el análisis de la hidrostática de la zona flotante, con el ánimo de determinar los límites de estabilidad de un puente líquido entre discos de igual diámetro. La estabilidad de una zona flotante cilíndrica fue ya tratada por Rayleigh en el siglo pasado en el análisis de chorros capilares; sin embargo, toda similitud entre chorros capilares y zona flotante acaba en este caso. El fin de los estudios estáticos es establecer qué formas de equilibrio puede adoptar el puente líquido al variar el volumen de líquido comprendido entre los discos y la distancia entre éstos, y de entre estas formas cuáles son estables (que serán las que se presenten en la realidad) y cuáles inestables. El problema así planteado requiere la búsqueda de las superficies de revolución de curvatura constante capaces de cumplir las condiciones impuestas por los discos y, de entre todas ellas, las de mínima energía que sean accesibles desde unas condiciones realizables.

El problema de determinar las superficies de equilibrio no es difícil (para zonas axilsimétricas estas superficies son: el cilindro, la esfera, la catenoide, las onduloides y las nodoides), pudiéndose calcular la ecuación de la línea meridiana en función de las integrales elípticas de primera y segunda especie.



4. METODO de la zona flotante para la purificación de materiales de alto punto de fusión y para la fabricación de monocristales. La varilla de material a tratar se sujeta por sus extremos, calentándola localmente cerca de uno de ellos hasta conseguir la fusión de una porción de la varilla. A continuación se desplaza lentamente la fuente térmica a lo largo de la barra. Durante el proceso, la varilla se mantiene en rotación para uniformizar la temperatura en la parte fundida. Al final, toda la varilla se habrá fundido y solidificado y, como las impurezas tienden a permanecer en el fundido, el material solidificado será de mayor pureza que el primitivo, ya que las impurezas habrán sido arrastradas hacia uno de los extremos de la varilla. Asimismo, el proceso permite controlar el avance del frente de fusión y por tanto controlar más fácilmente el crecimiento cristalino. En la figura se identifican los diversos componentes de acuerdo con la siguiente clave: material recristalizado (*1*), material fundido (*2*), material por tratar (*3*) y fuente térmica (*4*). Las flechas indican movimientos relativos entre fuente térmica y varilla.

Pero el problema de la estabilidad requiere un tratamiento más sutil.

Un primer límite de estabilidad viene impuesto por la máxima capacidad de la zona: fijado un cierto valor de la esbeltez de la zona (la esbeltez se define como el cociente entre la distancia entre discos y el diámetro de los mismos), al aumentar el volumen del líquido aumenta también el ángulo de contacto en el borde y, de continuar inyectando líquido en la zona, llegará un momento en que se desborde, mojando la superficie lateral del disco. Se suele tomar el valor de 180 grados como el máximo

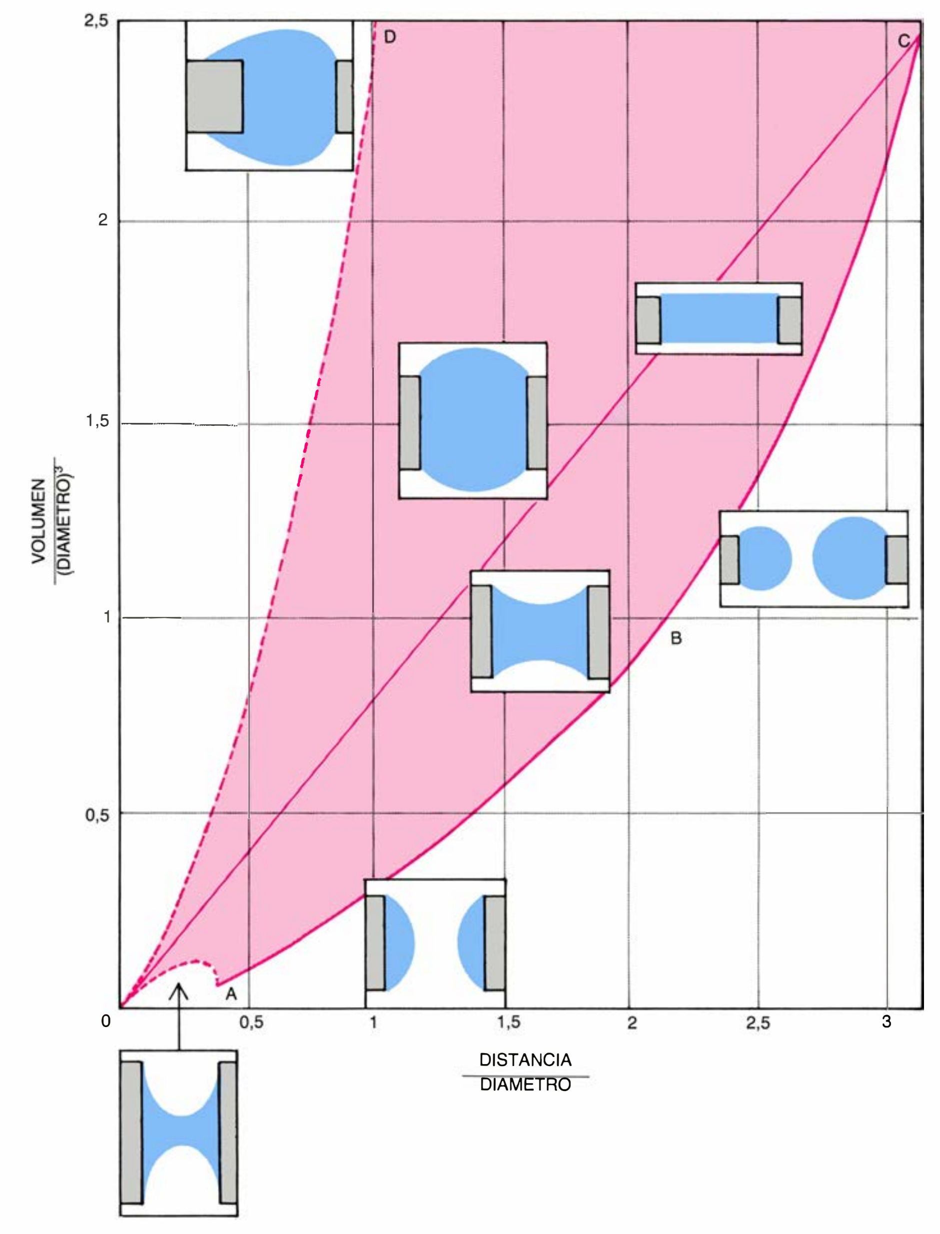
para este ángulo, y si se eligiera cualquier otro valor el volumen máximo variaría, pero el comportamiento sería el mismo. En cuanto a la limitación inferior (volumen mínimo) el caso difiere bastante, ya que una acotación basada en el valor del ángulo de contacto sólo es válida para valores muy pequeños de la esbeltez.

En el análisis del límite de estabilidad de mínimo volumen aparecen tres condicionantes que delimitan dicho límite, aplicándose uno u otro según sea la esbeltez del puente líquido. Para esbelteces muy pequeñas, el límite de volumen mínimo viene fijado por el desprendimiento del líquido del borde de los discos. Este límite depende del valor mínimo del ángulo de contacto sólido-líquido-gas. Su significado es el siguiente: para cada valor de la esbeltez, existe un volumen mínimo capaz de mantener la zona anclada a los bordes de los discos; si llegado a este límite se disminuye aún más el volumen, la línea de mojado se desprende del borde, retrayéndose hacia el centro de los mismos.

Para puentes líquidos más esbeltos, el límite inferior no viene impuesto por este condicionante de no desprendimiento de la línea triple, sino que existe, para zonas cuya esbeltez oscila entre los valores 1 y 2 aproximadamente, un volumen mínimo real. Esto quiere decir que si imaginamos un experimento en el cual, manteniendo la esbeltez constante, se va extrayendo líquido de la zona con la lentitud apropiada de modo que los sucesivos estados puedan considerarse de equilibrio, al ir disminuyendo el volumen de líquido se pueden encontrar formas de equilibrio que satisfagan las condiciones dadas de esbeltez y volumen. Esto ocurre hasta que, para cada esbeltez, se alcanza cierta cota del volumen y rebasada ésta ya no existen formas de equilibrio compatibles con el volumen actual del puente que, por tanto, evolucionará hacia la rotura. Las formas de equilibrio antes del límite de rotura son simétricas respecto al plano medio paralelo a los discos, el cuello de la zona coincide con este plano medio y se mantiene en esta posición durante el proceso de rotura. Así, para estos valores de la esbeltez, la rotura presenta un carácter simétrico, y la configuración final es dos gotas de igual volumen unidas una a cada disco.

Si la esbeltez es mayor que 2 (formalmente, mayor que 2,13) aparece un condicionante más severo aún que el de la forma estable de volumen mínimo. En efecto, antes de que se alcance esta forma de volumen mínimo tiene lugar una bifurcación hacia formas no simétricas (llamadas formas en ánfora) que originan una rotura no simétrica, con una configuración final consistente en dos gotas, de volúmenes diferentes, adheridas a los discos. El carácter de esta bifurcación queda patente recurriendo a una pequeña ayuda gráfica. Refiriéndonos a puentes líquidos cuyo volumen sea igual al de un cilindro comprendido entre los discos, podemos representar su energía superficial en función de la esbeltez y de la forma. (Piénsese que para un mismo volumen existen diversas formas de equilibrio, unas estables y otras inestables: cilindro, casquetes esféricos, formas en ánfora, etc., y cada una de estas geometrías puede individualizarse por el valor de un parámetro, por ejemplo, el ángulo de contacto en el borde de los discos.)

Ante cualquier perturbación, el puente líquido evolucionará hacia las configuraciones de mínima energía, análogamente a como se desplazaría una canica sobre esta superficie a lo largo de las curvas de esbeltez constante. Los valles y cimas de estas curvas corresponden a las formas estables e inestables, respectivamente, y el resto de los puntos a configuraciones que no serían de equilibrio. Para esbelteces in-



5. DIAGRAMAS DE ESTABILIDAD de una zona flotante mantenida entre discos de diámetro D, separados una distancia L, con un volumen de líquido V. Los esquemas representan las configuraciones existentes en cada región. El límite de estabilidad por desprendimiento del borde de los discos de la línea de mojado, línea *OA*, se ha calculado suponiendo un ángulo de contacto límite nulo. *AB* corresponde a la línea de volumen mínimo, que delimita la región de rotura simétrica. La línea *BC* corresponde a la rotura por bifurcación hacia formas no simétricas (rotura asimétrica). El límite superior, *OD*, se ha determinado admitiendo un ángulo de mojado máximo de 180°, por encima del cual se produciría el desbordamiento. En cada caso, el valor del ángulo máximo depende de los materiales. La línea *OC* representa la sucesión de formas cilíndricas.

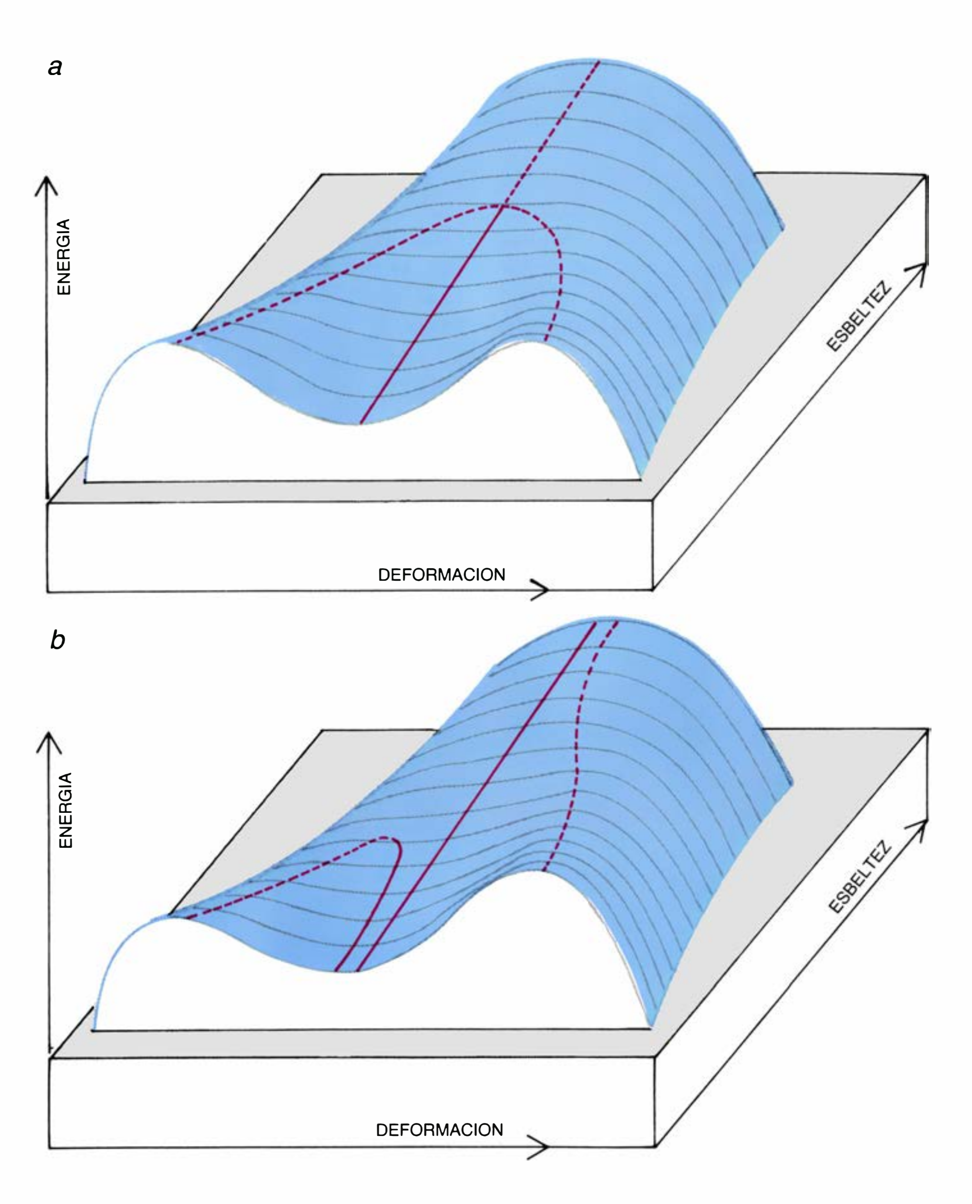
feriores a π, el cilindro es una configuración de equilibrio estable, ya que ante cualquier perturbación que modifique levemente la forma de la zona, la evolución será tal que restablezca la forma cilíndrica primitiva, siempre que la perturbación no obligue a sobrepasar las cimas que canalizan el valle de las formas cilíndricas (cimas que representan la energía de las formas en ánfora), pues si esto ocurriera la zona evolucionaría hacia otra configuración de menor energía aún, produciéndose la rotura.

Sin embargo, la energía de las formas cilíndricas aumenta linealmente con la esbeltez, mientras que en las formas en ánfora el crecimiento es más lento, resultando por tanto cada vez menor la perturbación mínima necesaria para romper la columna. Así se llega a un punto, cuando la esbeltez vale π, en el que el valle desaparece, y en esta posición, también de equilibrio, la más pequeña perturbación significará una evolución hacia la rotura: el equilibrio es inestable.

Lo expuesto hasta ahora se refiere a una configuración básica: puente líquido entre discos iguales en ausencia de fuerzas gravitatorias, que era la configuración prevista inicialmente para el experimento a realizar a bordo del Spacelab. Sin embargo, durante la etapa final de definición del experimento se acordó que, por razones de utillaje, los discos de trabajo no fueran exactamente iguales, lo que llevó a incluir discos desiguales, así como una pequeña gravedad axial, en el análisis de la estabilidad de puentes líquidos esbeltos. Cualquiera de estas perturbaciones ocasiona una pérdida de estabilidad: si existe gravedad, para que el puente líquido no rompa se debe disminuir la separación entre los discos o, si se mantiene ésta, aumentar el volumen de líquido. Lo mismo ocurre si los discos son desiguales. La razón de esta pérdida de estabilidad está en el carácter no simétrico de las perturbaciones consideradas que, para zonas largas, fuerzan el desarrollo de los modos no simétricos (inestables) de deformación de la entrefase. Cualquiera de estos efectos introduce un nuevo término en el conjunto de la energía del puente líquido que modifica el diagrama energético de un modo asimétrico: una de las cumbres aumenta su altura relativa mientras que la otra disminuye conectando con el valle de energías mínimas.

Desde el punto de vista práctico, lo más interesante es que, aunque estas perturbaciones son desestabilizantes, cada una de ellas también puede tener un efecto estabilizador frente a la otra. Supongamos una columna líquida mantenida verticalmente entre dos discos y sometida a un cierto nivel de gravedad; el líquido tenderá a desplazarse hacia el disco inferior, apareciendo un cuello cerca del disco superior. Si los discos son desiguales y el de menor diámetro se coloca arriba, estaremos acortando el radio del puente líquido justo en la región donde aparece el cuello provocado por la gravedad, forzando más aún la deformación de la entrefase. Lo contrario ocurre cuando el disco mayor se coloca arriba; las deformaciones de la entrefase provocadas por la desigualdad de los discos se oponen a las producidas por la gravedad.

Si los orígenes de los estudios de la hidrostática del puente líquido tienen ya más de un siglo, la mayoría de los efectos dinámicos se han abordado en los últimos cinco años. Se ha analizado, fundamentalmente, el proceso de ro-

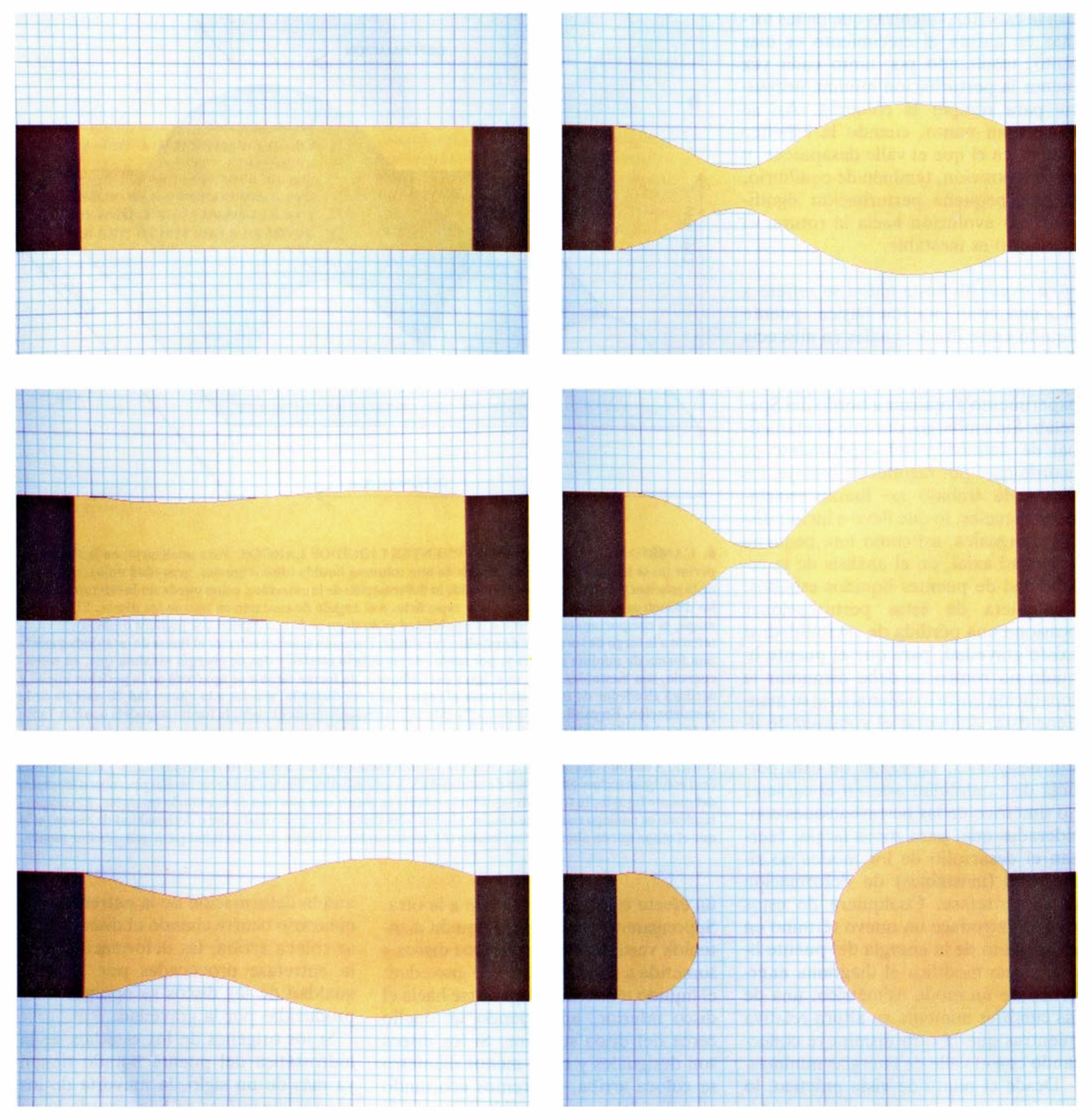


6. CAMBIO DE ESTABILIDAD DE PUENTES LIQUIDOS LARGOS. Para analizarlo, en la figura superior (*a*) se ha representado la energía de una columna líquida (discos iguales, gravedad nula) en función de la esbeltez y de un parámetro que mide la deformación de la entrefase, como puede ser la diferencia entre los diámetros máximo y mínimo de la superficie, o el ángulo de contacto en uno de los discos. Fijado este mapa, el comportamiento frente a la estabilidad es análogo al de una canica que pudiera deslizar sin rozamiento a la largo de las curvas de esbeltez constante. Para una esbeltez inferior al límite de estabilidad existe una forma de mínima energía y, por tanto, de equilibrio estable y dos de energía máxima que, en consecuencia, son inestables. Ante una ligera perturbación, la forma estable se apartará de la posición inicial y oscilará alrededor de ésta hasta volver al reposo una vez disipada, gracias a la viscosidad, la energía de la perturbación. En cambio, si la perturbación es lo suficientemente grande, la forma se podría apartar tanto de la posición de equilibrio estable que se podría rebasar la barrera de las cumbres. En el símil de la canica, ésta caería por una de las laderas exteriores, que en el puente líquido se traduce en una evolución hacia la rotura. A medida que la esbeltez aumenta, la diferencia entre las energías de la forma estable y las inestables decrece, igualándose en el límite de estabilidad. A esta esbeltez, cualquier perturbación provocará la rotura del puente líquido; la forma de equilibrio estable es ahora una posición de equilibrio inestable. El comportamiento es análogo cuando se añaden efectos tales como microgravedad axial o desigualdad de discos, sólo que el mapa queda ahora distorsionado (*b*), alcanzándose el límite de estabilidad para una esbeltez menor.

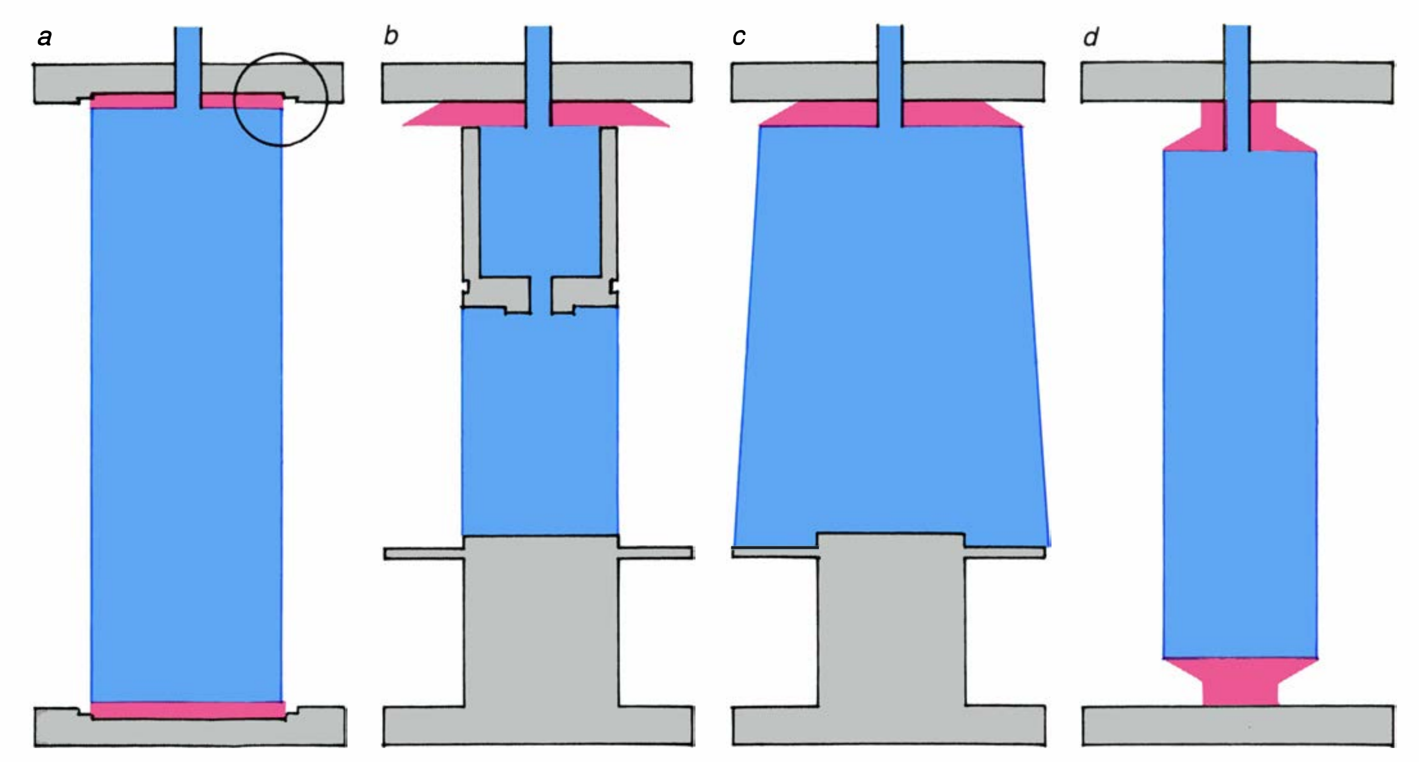
tura del puente líquido. Para ello se aplicó al problema del puente líquido un modelo unidimensional de "rodajas" no viscoso, usado antes en el estudio de la rotura de chorros capilares, y un modelo de Cosserat desarrollado originalmente en Elasticidad. Ambos modelos proporcinan parecidos resultados, aunque el de Cosserat incluye efectos de inercia transversal y también efectos viscosos.

La mayoría de los estudios dinámicos se ha realizado numéricamente, analizándose la evolución de puentes líquidos con muy diversas configuraciones y sometidos a diferentes perturbaciones. La consecuencia más importante del análisis dinámico, aparte de predecir la variación con el tiempo de la forma de la superficie libre, campos de velocidades y presiones, períodos de vibración, tiempos de rotura y otros detalles, es que permite estimar características globales del movimiento susceptibles de comprobación experimental; tal es el caso de los volúmenes de las gotas en que se divide la zona cuando rompe, cuyo valor depende de parámetros controlables (la separación entre los discos y el volumen de líquido), pero no, afortunadamente para el experimentador, de la naturaleza de la perturbación que desencadena la rotura. Es así porque la rotura constituye un fenómeno gobernado por la tensión superficial.

También se ha analizado desde el punto de vista dinámico la influencia de un medio exterior de densidad comparable a la del puente líquido con el propósito de determinar qué parámetros de la evolución resultan afectados por la existencia de este medio. Con estos estudios se pretende hallar



7. ROTURA DE UN PUENTE LIQUIDO ESBELTO que se halla próximo al límite de estabilidad. Los resultados presentados en esta sucesión de diagramas se han obtenido mediante simulación en ordenador: modelo unidimensional de la zona flotante. (La rotura está gobernada por la tensión superficial.)



8. LA CONFIGURACIÓN NOMINAL prevista para los ensayos a bordo del *Spacelab 1* (*a*) resultó inalcanzable debido al desbordamiento del líquido de trabajo. La ranura del borde de los discos no logró contener el avance de la línea triple a pesar de que geometrías análogas se habían ensayado con éxito en tierra en condiciones de ingravidez simulada mediante flotabilidad neutra. La razón de tan diferente comportamiento está, probablemente, en el medio que rodea al líquido de trabajo: aire en los experimentos en el espacio y otro fluido de la misma densidad en la técnica de flotabilidad neutra. Gracias al entusiasmo tanto de los astronautas, B. Lichtemberg y U. Merbold, como de nuestro colega desplazado al Centro de Control de Operaciones en Houston, I. Martínez, se pudo, con los materiales disponibles a bordo, sustituir los discos de trabajo por otros y realizar la secuencia experimental con puentes líquidos entre discos iguales (*b*) y desiguales (*c*). Esta experiencia aconsejó variar la geometría de los discos para el experimento de la misión *Spacelab D1*, en la que los discos fueran cónicos (*d*).

un apoyo teórico que permita realizar experimentos en microgravedad simulada mediante la técnica de la flotabilidad neutra, lo que reduciría la necesidad de experimentación en el espacio. Los resultados obtenidos indican la viabilidad de esa línea de trabajo para configuraciones axilsimétricas, lo que ha motivado el inicio de un programa de experimentación en tierra.

Otro efecto dinámico estudiado es el de discos en isorrotación con una pequeña gravedad axial. Se han analizado los dos modos fundamentales de deformación de la entrefase: el modo axilsimétrico y el modo *C* o en comba (observado por primera vez en los experimentos realizados en el *Skylab 4,* en 1973).

Una parte importante del esfuerzo asignado en el Laboratorio de Aerodinámica de Madrid al estudio del comportamiento de puentes líquidos está dedicado a la verificación experimental en sus dos vertientes: experimentación en el espacio y experimentación en tierra. Nuestro grupo participó en la misión *Spacelab 1* con el experimento 1-ES-331 "Floating Zone Stability" (según la nomenclatura de la Agencia Espacial Europea).

El experimento 1-ES-331 fue propuesto por Ignacio Da Riva en 1974 con el fin de estudiar la estabilidad de una columna líquida en gravedad reducida. La configuración nominal consistía en una masa líquida, dimetil silicona, situada entre dos discos, sometida a las siguientes perturbaciones: rotación de los discos, contrarrotación de los discos, vibración axial de uno de los discos, desalineación de los ejes de los discos, rotura cilíndrica (separación de los discos inyectando líquido de modo que la zona permanezca cilíndrica hasta alcanzar el límite de estabilidad), rotura de la zona por estirado (separación de los discos manteniendo el volumen de líquido constante) y rotura de la zona por extracción de líquido.

El 1-ES-331 constituyó, junto con otros cinco experimentos europeos, el área de física de fluidos para la primera misión Spacelab. Para la realización de estos experimentos se diseñó y construyó un equipo de experimentación conocido como FPM (siglas de Fluid Physics Module). El 1-ES-331 era el último experimento programado en la secuencia de experimentos a realizar en el FPM debido al alto riesgo de contaminación de la cámara de ensayos en alguna rotura incontrolada de la columna líquida. Este problema se presentó en el experimento anterior al nuestro al intentar formar una columna líquida: el líquido, aceite de silicona, sobrepasó el borde del disco de inyección, extendiéndose el aceite por parte de la cámara de ensayos.

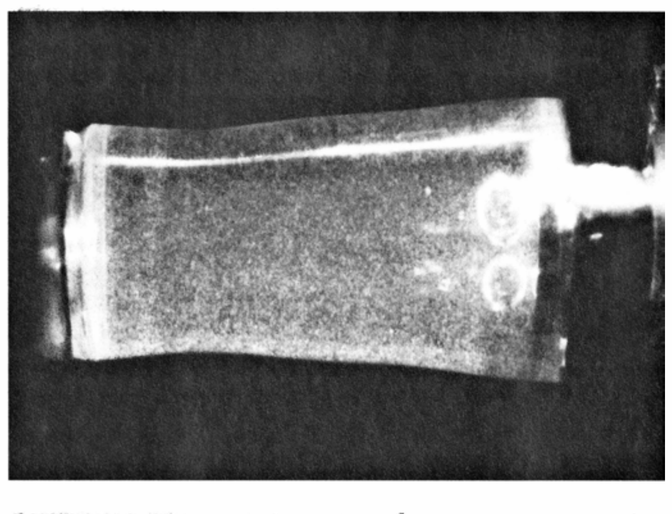
En nuestras previsiones para la realización del experimento, este problema estaba advertido en uno de los primeros puntos de la descripción de la secuencia en el que se aconsejaba "recoger, limpiar y volver a intentarlo". El desbordamiento se puede producir también en la simulación en tierra y, de hecho, los astronautas U. Merbold y B. Lichtemberg se habían familiarizado con este proceso durante las sesiones de entrenamiento.

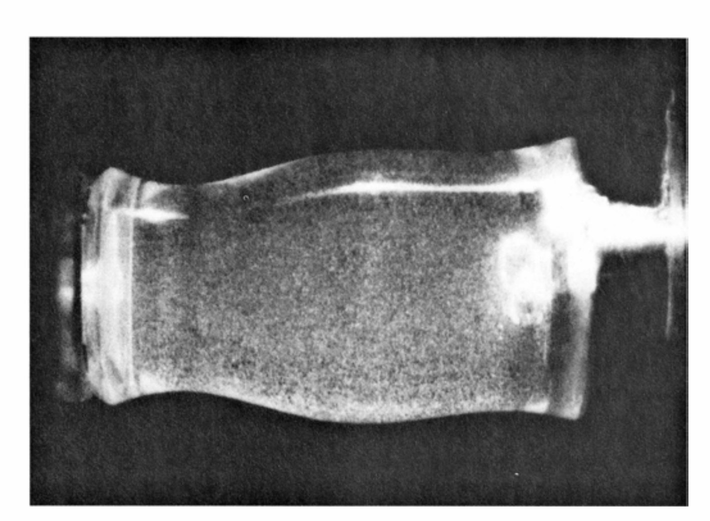
En nuestro experimento esta fatalidad volvió a repetirse: el líquido rebasó el borde del disco, lo que obligó a variar la secuencia nominal, concentrando el esfuerzo en una búsqueda sistemática de la causa de este comportamiento. Se identificó una posible interacción entre el proceso de avance de la línea de mojado con el anclaje en el borde del disco: a velocidades suficientemente bajas como para evitar el desprendimiento de líquido en forma de chorro en el orificio de inyección, se producía una onda de avance del líquido que sobrepasaba la barrera constituida por el pequeño borde, e incluso la corona exterior, recubierta con una sustancia antimigratoria.

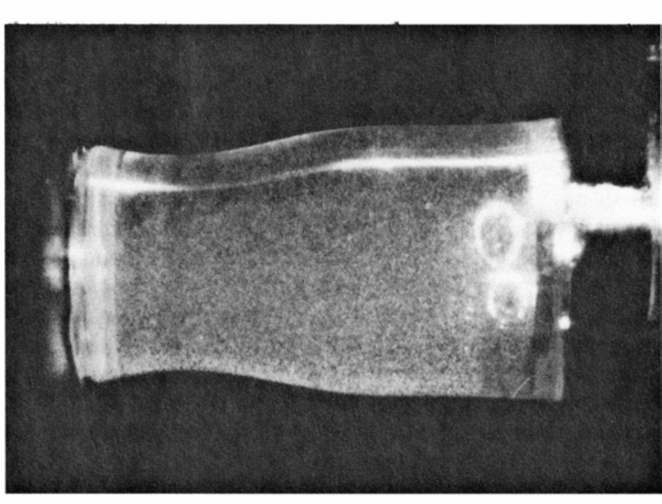
A continuación se realizaron ensayos

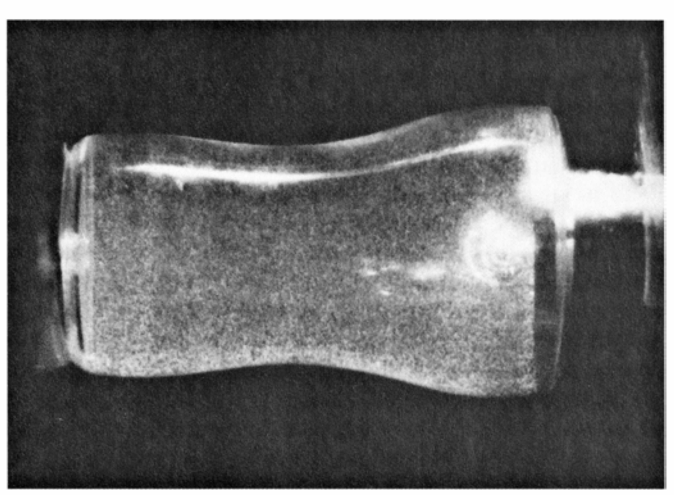
inyectando manualmente el líquido con gran cuidado, obteniéndose una gran gota de líquido anclada al borde del disco de inyección. Aumentando lentamente el volumen de la gota, se encontró un ángulo máximo en el borde entre 70 y 80 grados, más allá del cual el aceite se desbordaba rápidamente. A la velocidad nominal de inyección, la onda de avance de mojado producía el rebosamiento a ángulos más pequeños, entre 20 y 30 grados. Además, se observó que la barrera antimigratoria aumentaba este ángulo máximo en unos 10 grados.

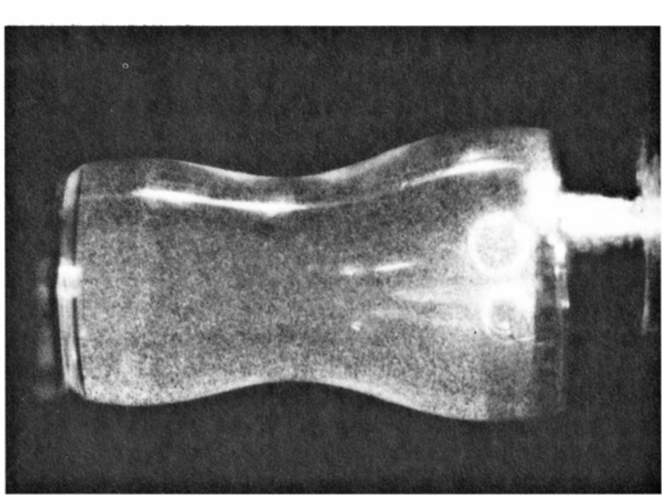
La segunda parte del tiempo asignado se empleó en profundizar en las anomalías encontradas y que impedían el progreso en la secuencia de experimentación prevista. Primero se investigó el límite máximo de velocidad de inyección para que no se produjera un chorro en el orificio de salida, problema que había sido previsto durante el diseño del FPM. Se encontró que dicho límite se encontraba entre 3,5 y 4 centímetros cúbicos por segundo, para un orificio de inyección de 6 milímetros de diámetro. En segundo lugar, se ensayó la posibilidad de establecer una zona por un camino diferente al de la formación de una gota: llenando el espacio entre los discos antes situados a corta distancia (7,5 mm). Se obtuvo un puente líquido que se consiguió hacer crecer hasta 10 mm, distancia a la que

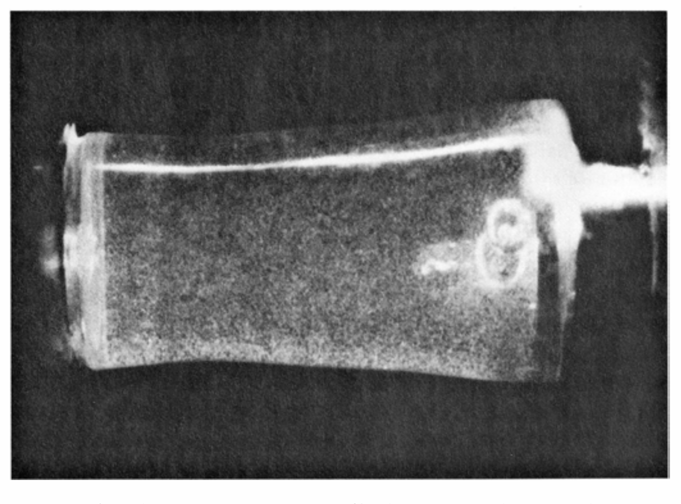












9. ENSAYOS DE VIBRACION del puente líquido a bordo del Spacelab. En las fotografías se muestran seis fotogramas consecutivos tomados con un intervalo de tiempo de 1 segundo. El disco menor, de 51 milímetros de diámetro, permanecía fijo, mientras que el mayor, de 60 milímetros de diámetro, vibraba en la secuencia presentada con una frecuencia de 0,4 hertz y una amplitud de 0,5 mm. Dentro del puente líquido se observan trazadores y dos burbujas de aire.

se produjo el desbordamiento. Por último, se analizó la posible influencia, en el comportamiento del líquido, del hecho de que los discos fueran de diferentes materiales, uno de aluminio y otro de acero inoxidable; para ello se pusieron en rotación, una vez formada la zona, no encontrándose diferencias apreciables. El cambio de la viscosidad del líquido de trabajo no produjo ninguna mejora.

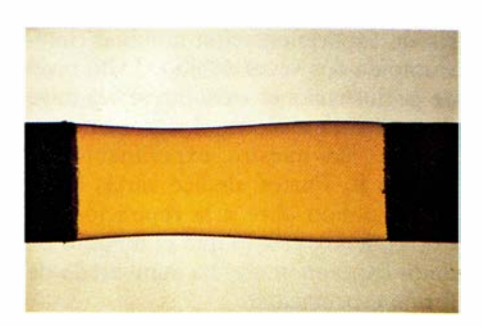
Así estaban las cosas cuando las autoridades de la ESA y la NASA decidieron prolongar un día la duración de la misión. De nuestro equipo, se había desplazado Isidoro Martínez al Centro de Control de Operaciones en Houston, por lo que recayó sobre él (la diferencia horaria impidió la comunicación con Madrid) la decisión de cómo usar el tiempo extra asignado y, entre las posibilidades que se brindaban para continuar con el experimento, cambiar los líquidos a emplear o los discos de trabajo, y tomó la decisión de realizar esta última. Se cursaron instrucciones a los astronautas para cambiar los discos, tomando piezas hasta el momento empleadas en otro de los experimentos del FPM. Debido al ruido presente en algunos momentos en la línea de comunicación, las instrucciones no llegaron claramente a la nave y, además, las dimensiones reales del disco seleccionado no coincidían con las especificadas en la documentación existente en tierra. Todo ello originó un malentendido, que se solucionó gracias a la capacidad de iniciativa de Lichtemberg, quien, consciente de los fines científicos del experimento, instaló una pieza de igual diámetro que el disco opuesto, aunque distinta de la seleccionada por el Dr. Isidoro Martínez.

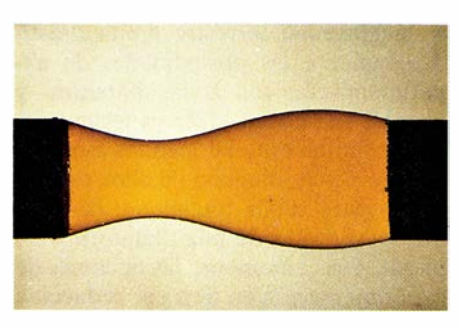
Con esta configuración se formó una columna cilíndrica de 30 mm de diámetro y la máxima longitud posible, 52 mm, y se pusieron ambos discos en isorrotación. Para seguir el movimiento en el interior de la zona, el líquido contenía unas minúsculas partículas trazadoras. La existencia de unas pequeñas burbujas de aire introducidas accidentalmente ayudaron a detectar con facilidad el movimiento interno en las secuencias grabadas en TV. Se fue aumentando la velocidad de giro con el fin de detectar la existencia de ciertas inestabilidades, y a 15 revoluciones por minuto (rpm) se observó el comienzo de la deformación de la zona según el modo *C*, aunque en forma estable. La teoría predice que dicha inestabilidad se debe producir a 21 rpm; una posible explicación de esta anticipación es la existencia de un umbral en la velocidad de rotación asociado con la desalineación de los discos (unos 2 mm en este caso debido a problemas de montaje de las piezas cambiadas) que dispare la deformación en *C*, cuya amplitud crecería con la velocidad de rotación hasta llegar a la rotura al alcanzar un límite próximo a 21 rpm (valor para el caso sin excentricidad).

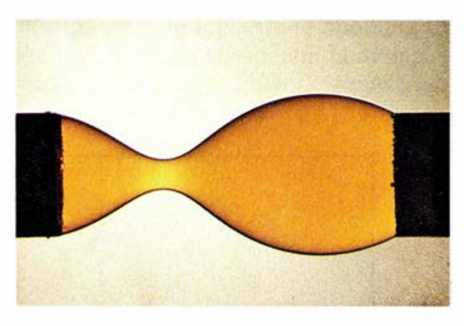
Debe mencionarse que, mientras tanto, el líquido iba saliendo lentamente de la columna líquida a través de las paredes laterales del soporte improvisado (una pieza roscada). Esta pérdida fue detectada al cabo de un rato al observar un pequeño estrechamiento de la parte central de la columna. Se añadió líquido para recuperar la forma cilíndrica y volvió a repetirse la secuencia de rotación con incrementos de velocidad más pequeños; pero, cuando la rotación en modo *C* se desarrolló a 12 rpm, aumentó la pérdida de líquido. A 14 rpm era ya bastante notoria, produciéndose una rotura lenta, simétrica como era de esperar para el valor de la esbeltez de la columna. Se formó una gota en el disco de inyección con un ángulo de mojado en el borde próximo a 90 grados, pero el retroceso de la otra gota hacia el disco opuesto produjo el desprendimiento de la línea triple del borde de 30 mm de diámetro, sobrepasándolo hasta alcanzar el borde exterior (60 mm de diámetro).

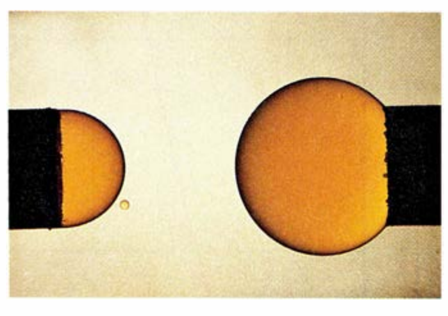
Cuando se tuvo contacto visual pudo aclararse el malentendido anterior, y se montó el disco seleccionado que, aunque de diferente diámetro, tenía un borde agudo. Con la nueva configuración llegaron a formarse zonas de 96 mm de longitud, ensayándose en ellas casi todas las perturbaciones programadas para el experimento 331: excitación de la vibración axial, obtención de la deformación en modo *C*, roturas, etcétera.

Así pues, el resultado global del experimento 1-ES-331 puede considerarse como satisfactorio, a pesar del gran número de dificultades encontradas, cuya causa (la ausencia de los conocimientos adecuados que permitan un claro manejo de los líquidos en condiciones de gravedad reducida) es precisamente el objeto de este y otros muchos experimentos. La experiencia adquirida durante este primer vuelo nos ha servido para diseñar experimentos posteriores evitando los problemas relativos al manejo de líquidos en ingravidez. Así, a finales de 1984 se realizaron ensayos a bordo de aviones KC-135, dentro de un programa de vuelos parabólicos patrocinado por la Agencia Espacial Europea, y más recientemente se han realizado otros dos experimentos. El primero, en mayo de 1985, a bordo de un cohete de sondeo del programa alemán TEXUS, realizándose de un modo totalmente automático diversas zonas líquidas esbeltas, variando de un caso a otro la velocidad de inyección del líquido. El segundo experimento ha tenido lugar durante la segunda misión europea del laboratorio espacial Spacelab (*Misión D1*) en noviembre de 1985 y ha consistido, básicamente, en la repetición del experimento 1-ES-331, con notable éxito esta vez, pues, aparte de realizar la secuen-









10. SECUENCIA DE ROTURA de un puente líquido esbelto. El puente se ha formado con un aceite mineral (dimetil silicona, tintada con anilina, con una viscosidad de 20 centistokes) sumergido en un baño de alcohol y agua. Los discos son de metacrilato, de 30 milímetros de diámetro, pintados de negro con una pintura de características oleófilas. En la superficie lateral de los discos esta pintura está a su vez recubierta por otra transparente oleófoba.

cia de experimentación nominal (interrumpida dos veces debido al alto nivel de perturbaciones existente en la nave durante esta misión), el astronauta encargado de nuestro experimento, el alemán R. Furrer, dedicó varias horas de su tiempo libre a la repetición del experimento, con lo que el número de datos experimentales ha aumentado de forma espectacular.

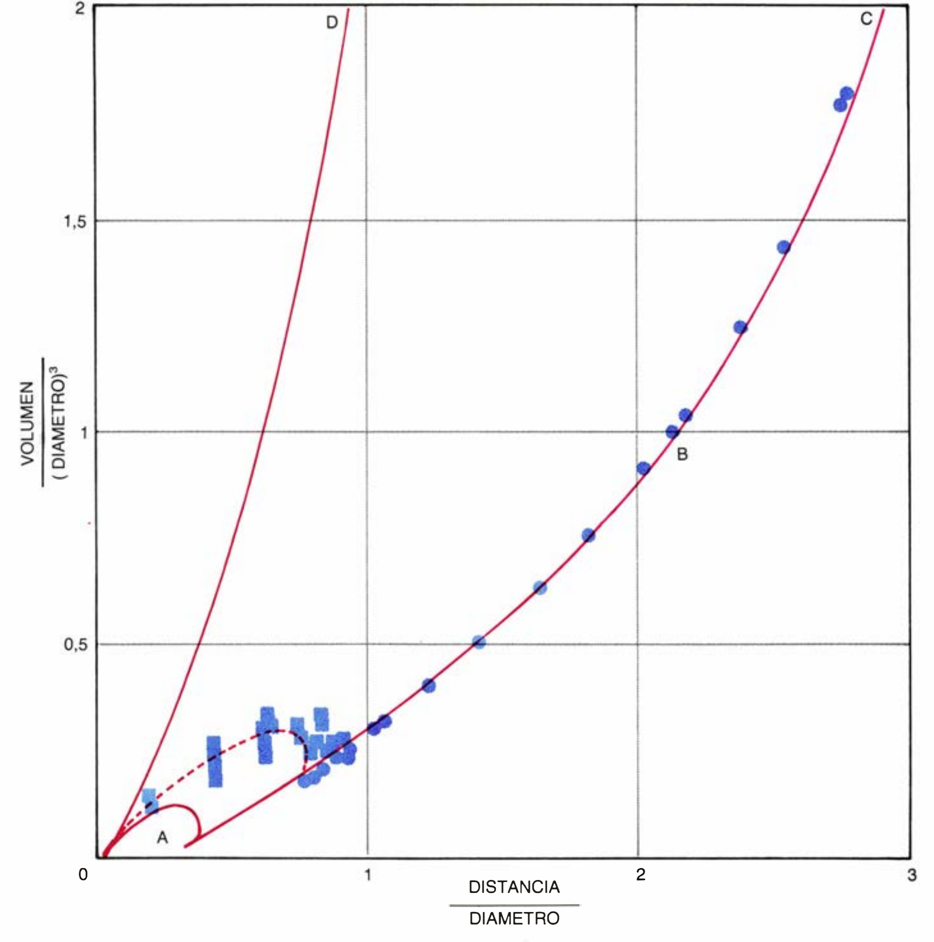
La gravedad terrestre limita drásticamente las posibilidades de experimentación con zonas flotantes, y aunque se pueden idear experimentos que permitan estudiar algunos aspectos del problema, ninguno de ellos es adecuado para suplir totalmente las condiciones ideales de una plataforma espacial. Por el momento las opciones de experimentación en tierra se reducen a dos: zonas submilimétricas y baños de flotabilidad neutra. La primera opción requiere el manejo de puentes líquidos de dimensiones diminutas. La dificultad de esta opción reside precisamente en el tamaño de la zona, que necesita del uso de un equipo de extrema precisión.

La otra opción reside en la utilización de baños de flotabilidad neutra o baños de Plateau. Esta técnica consiste en sumergir un líquido de densidad conocida en un baño de otro líquido, inmiscible con el primero, que tenga la misma densidad; de este modo, las fuerzas gravitatorias y la flotabilidad se cancelan entre sí y la entrefase entre los dos líquidos adopta la misma forma que la que tendría en condiciones de gravedad nula. La mayor ventaja de este método es que no impone restricciones a las dimensiones de la zona (salvo las económicas derivadas del coste de los líquidos de ensayo). Entre las desventajas cabe destacar, en primer lugar, que aunque la técnica es particularmente apta para ensayos estáticos puede resultar deficiente para los dinámicos (es preciso evaluar las modificaciones que introduce el baño exterior, lo que hasta ahora sólo se ha hecho para el caso de evoluciones estrictamente axilsimétricas) y es inadecuada para ensayos térmicos (debido a las corrientes de convección libre que no aparecen en condiciones de gravedad nula).

La preparación de un experimento en baño de Plateau requiere el concurso de tres líquidos de distintas densidades, con el líquido de densidad intermedia inmiscible con la mezcla de los otros dos. La terna más empleada en nuestro laboratorio es aquella que utiliza dimetil silicona como líquido de trabajo y una mezcla de agua destilada y alcohol metílico como baño exterior, y en menor grado la terna formada por una mezcla de dibutilftalato y dioctilftalato para líquido de trabajo y agua destilada como baño.

Para este programa de experimentación se ha construido en el Laboratorio de Aerodinámica (además de otros recipientes de menores prestaciones) un equipo para el ensayo de zonas líquidas en tanque de Plateau denominado PTF (Plateau Tank Facility), diseñado de tal manera que los discos de trabajo puedan experimentar los mismos movimientos que en el FPM.

La utilidad de este equipo, aun dentro de las limitaciones impuestas por el medio exterior, ha sido grande ya que ha permitido trabajar en condiciones de fácil accesibilidad, riguroso control y sin las limitaciones de tiempo de ensayo existentes en la experimentación espacial. Dentro del plan de comprobación de resultados teóricos se han llevado a cabo diversos ensayos, tanto estáticos como dinámicos, y como botón de muestra valga la comprobación experimental de los límites de estabilidad estática en los casos de rotura y desprendimiento. La concordancia entre los resultados teóricos y experimentales es perfecta en el caso de rotura y, como era de esperar, menor en el caso de desprendimiento. En efecto, este segundo límite depende del valor del ángulo de contacto de la línea triple zona-baño-disco; este valor del ángulo de contacto está influido a su vez por las pequeñas imperfecciones existentes en los bordes de los discos debidas al mecanizado y al tratamiento superficial de los mismos, por la posible contaminación de la entrefase, etc., aspectos todos de muy difícil control. No obstante, los resultados obtenidos en esta parte del diagrama muestran la misma tendencia que las predicciones teóricas, con un ángulo de desprendimiento comprendido entre 30 y 40 grados.



11. COMPROBACION EXPERIMENTAL mediante la técnica de flotabilidad neutra de los límites de estabilidad de mínimo volumen. Es de destacar la absoluta concordancia entre los resultados teóricos y experimentales cuando se alcanzan situaciones de rotura. (Los resultados experimentales se denotan mediante círculos para estos casos.) La razón de esta concordancia estriba en que son muy reproducibles en la experimentación las hipótesis realizadas en el estudio estático. En el caso de desprendimiento del borde de los discos, el comportamiento es más disperso ya que éste se produce cuando se alcanza el ángulo de contacto mínimo, ángulo que depende de factores difícilmente controlables como son irregularidades en el acabado superficial de los discos, contaminación de las entrefases, etcétera. A pesar de todo, se puede observar un agrupamiento de los resultados experimentales (*representados por un cuadrado*) en torno al límite de desprendimiento teórico correspondiente a un ángulo de contacto mínimo de 40 grados (*línea punteada*).

Juegos de ordenador

Exploración de los algoritmos genéticos en un mar primordial lleno de "glovitos"

A. K. Dewdney

Imaginemos un mar abstracto, habitado por organismos abstractos llamados "glóbulos vivos finitos", a los que por eufonía dejaremos en "glovitos". Cada glovito está dotado del más sencillo aparato de decisión posible. Los glovitos serían equivalentes biológicos de lo que en ciencias de cómputo se denominan autómatas finitos. Cada glovito contiene también un único cromosoma, consistente en una ristra de símbolos en los que está codificado el autómata. Los glovitos habitan en un caldo digital –el caldo primordial– que se encuentra en constante fluir. Para que el glovito sobreviva es preciso que logre predecir con exactitud los cambios que vayan a producirse en su ambiente.

En la sopa primordial que puse recientemente a fuego lento en mi ordenador, los glovitos que no acertaban a atinar en sus predicciones se extinguían. Los más acertados dejaban progenie, la cual lograba a veces mejorar el índice de acierto de sus ancestros. Al cabo, terminó por producirse evolutivamente una estirpe de individuos capaces de predicción perfecta.

Los glovitos y sus tendencias evolutivas ilustran con gran pulcritud una forma de programación conocida como "algoritmo genético". Esta técnica, de la que fue pionero hará unos veinte años John H. Holland, de la Universidad de Michigan, permite en ocasiones resolver difíciles problemas, por obtención evolutiva de una sucesión de soluciones aproximadas. Emparejando unas con otras las mejores de las antiguas soluciones se obtienen soluciones aproximadas nuevas. No tarda mucho en aparecer una solución mejor que sus progenitoras, que pasa a encuadrarse entre las reproductoras preferentes. Se han aplicado con cierto éxito algoritmos genéticos al reconocimiento de formas y patrones, a sistemas clasificadores, al funcionamiento de oleoductos, al diseño y distribución de símbolos y a unos pocos problemas más. En mi sopa computarizada, esta técnica produjo una estirpe de glovitos de comportamiento superior. ¿Se debió tal éxito a la eficacia general del método de algoritmo genético, o a la simplicidad de la tarea de predicción que debían afrontar los glovitos? Resulta difícil responder a esta pregunta. Los lectores interesados pueden reflexionar sobre ella y, si tienen a su alcance un ordenador, reproducir el fenómeno subyacente.

Un autómata finito puede encontrarse en un número finito de estados; la recepción de una señal de entrada lo hace pasar automáticamente de un estado a otro. La clase de autómatas utilizados en los glovitos también generan señales. En el seno del autómata, las señales recibidas se representan por símbolos. Cuando se recibe una señal, el autómata cambia de estado y emite una segunda señal.

Para representar el proceso resulta adecuada una tabla de transiciones de estados. Por ejemplo, un autómata que pueda adoptar tres estados, *A*, *B* y *C*, y sea capaz de manejar unos y ceros aferentes y eferentes, queda pulcramente representado mediante una tabla de 3 por 4. Para cada estado en que el autómata pueda encontrarse, y para cada símbolo que pueda recibir, la tabla dispone de dos entradas. La primera da el correspondiente símbolo de salida; la segunda, el estado que el autómata va a adoptar a continuación:

	0		1	
A	1	*B*	1	*C*
B	0	*C*	0	*B*
C	1	*A*	0	*A*

El autómata representado por esta tabla podría muy bien encontrarse en algún instante en el estado *C*. Si el autómata recibiera entonces un 1, la tabla nos diría que el autómata generaría un 0 y pasaría a adoptar el estado *A*.

Otra representación, que los humanos encuentran más fácil de leer, es el diagrama sagital de transición de estados. En él, los círculos representan estados y las flechas transiciones. Para denotar que un autómata pasa de un estado a otro cuando recibe un símbolo específico, se traza una flecha que vaya de uno a otro. La flecha debe ir rotulada tanto con el símbolo de entrada, causante de la transición, como con el símbolo de salida resultante [*véase la figura 1*].

Los autómatas finitos comienzan a funcionar partiendo siempre de un estado específico, llamado inicial. Cada vez que un imaginario reloj da un golpe de péndulo, llega un nuevo símbolo, se emite otro nuevo símbolo y se adopta un nuevo estado. Los autómatas que he utilizado para mis glovitos reciben y envían todos dos únicos símbolos, siempre los mismos: 0 y 1.

¿Cómo ha de interpretarse el comportamiento de los glovitos, si tan poco se conoce de la biología de esas criaturas? En ello precisamente reside el placer de la abstracción. Los símbolos recibidos por el autómata son meros mensajes sensorios procedentes del entorno. Correlativamente, un símbolo de salida puede entenderse por la respuesta que el organismo da al estado más reciente de su medio ambiente.

Tan versátil es la noción de glovito, que sus entradas y salidas pueden representar gran diversidad de fenómenos biológicos específicos. Por ejemplo, una señal de entrada podría representar un gradiente térmico o químico. El correspondiente símbolo de salida podría ser una orden dirigida a un efector que controlase las vibraciones de cilios, o que tuviera a su cargo un mecanismo de esporulación. Una tarea que reviste gran importancia para todo ser vivo que desee evolucionar hasta un nivel mínimamente aceptable (catedrático de universidad, pongamos por caso) es la predicción de los cambios de su entorno. Para los glovitos, su ambiente vital consiste en una secuencia, aparentemente interminable, de signos 0 y 1. En la medida en que los símbolos recibidos sean indicativos de acontecimientos importantes, comportará sin duda cierta ventaja que el glovito tenga capacidad de predecir cuál será el próximo símbolo, y tanto más si en alguna interpretación específica del funcionamiento del glovito se vieran reforzadas sus propias posibilidades de supervivencia.

La mayoría de los glovitos son bastante ineptos en lo tocante a predecir el comportamiento de su entorno. Por ejemplo, el glovito descrito por la tabla de transiciones dada responde a la se-

cuencia de señales ambientales

0 1 1 1 0 0 0 0 1 0 1 1 0 . . .

con las salidas

1 0 0 0 0 1 1 0 0 1 0 0 0 . . .

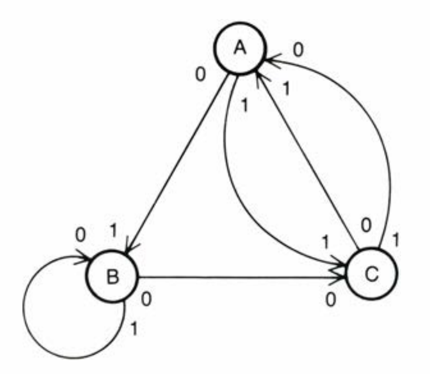
En cada etapa de su funcionamiento, la salida del glovito es su pronóstico de cuál será la próxima señal que llegará procedente del entorno. Para hallar el número de pronósticos correctos, se desplaza la secuencia de salida un símbolo hacia la derecha, y se la compara, bit a bit, con la secuencia de entrada. Contemos el número de símbolos concordantes. En este caso, el glovito tan sólo predijo correctamente seis de los doce símbolos que le fueron llegando, puntuación no superior a la que sería de esperar pronosticando al puro azar.

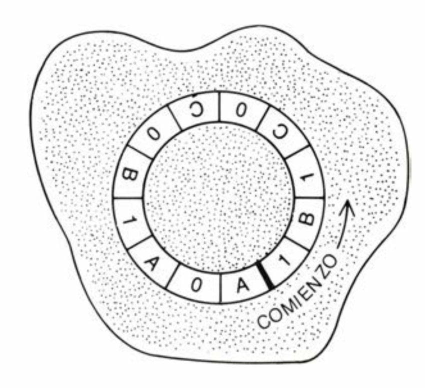
Es fácil pedir demasiado a los autómatas finitos. En efecto, no hay derecho a exigirle a un glovito que sea capaz de pronosticar correctamente las señales de medios ambientales no periódicos. Los lectores querrán, seguramente, reflexionar por un momento sobre esta cuestión. ¿Por qué habrá de constar toda secuencia correctamente predicha de una ristra fija de símbolos, interminablemente repetida? Por ejemplo, el glovito de tres estados que antes fracasó en la prueba de predicción que le propusimos, triunfa brillantemente en la siguiente secuencia ambiental:

0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 1 1 . . .

Pero aquí la sucesión de valores ambientales desfila al compás de un sencillo redoble, a saber, 010011.

Existen varias docenas de glovitos triestado, pero tan sólo unos pocos son capaces de predecir correctamente esta secuencia. Los glovitos de más de tres estados con capacidad de predicción perfecta para una sucesión ambiental determinada son raros, y escasean tanto más cuanto mayor es el número de estados. La predictibilidad depende fuertemente del período de la sucesión. Si la serie fundamental de símbolos es suficientemente larga, ningún glovito de número prefijado, *n*, de estados conseguirá jamás llegar a pronosticarla exactamente. Existe, evidentemente, una relación entre el número de estados que puede adoptar un glovito y la longitud máxima del período que es capaz de pronosticar perfectamente. Seguramente los lectores gusten de descubrir por sí mismos tal relación. ¿Cuál es la máxima longitud de período que





*1. Un diagrama de transiciones de estados (***izquierda***) y el glovito correspondiente, con su cromosoma (***derecha***)*

puede pronosticar un glovito de *n* estados?

Los glovitos son algo más que autómatas finitos que se esfuerzan por predecir su entorno, pues tienen un cromosoma. Por algún procedimiento (desconocido todavía) los glovitos engendran periódicamente nuevos glovitos. Al examinar el cromosoma del glovito y ponerlo en relación con el autómata finito correspondiente, vemos de qué modo los genes heredados determinan el comportamiento de la progenie. Comencemos con la tabla de transiciones de estados y vayamos separando, una por una, las filas, de arriba abajo. Unamos las filas, pegando la cola de cada una con la cabeza de la siguiente y, finalmente, el comienzo y el fin de la tira así obtenida. El resultado será un cromosoma circular.

Antes de llevar a cabo esta última operación de empalme, el cromosoma de nuestro ejemplar triestado se nos presenta como una ristra de 12 genes:

1 B 1 C 0 C 0 B 1 A 0 A

En puridad, los símbolos de esta serie son alelos. Un alelo es una forma específica de gen, que se asienta en un locus determinado. Los genes pueden, por tanto, especificarse tanto por su nombre como por su emplazamiento, o locus. Así pues, el séptimo símbolo, contando desde la izquierda, controla el símbolo de salida de un glovito cuando éste se encuentra en estado *B* y se recibe un 1 procedente del entorno. Aquí, el locus es 7.

Hace poco preparé en mi ordenador personal una sopa primordial, con 10 glovitos tetraestado. No habían transcurrido 1000 de las unidades de tiempo, que llamo cronones, cuado ninguno de los glovitos originales permanecía con vida. Todos habían sido remplazados por otros de mayor habilidad de predicción. La pantalla de mi ordenador iba mostrando las puntuaciones máxima y mínina alcanzadas por la población existente en ese momento. La puntuación mínima fluctuaba fuertemente; la máxima, en cambio, fue trepando poco a poco [*véase la figura 2*]. Justamente cuando comenzaba yo a desesperar de que pudiera evolucionar un predictor perfecto apareció uno súbitamente y, a partir de ahí, la puntuación máxima quedó estancada en 100.

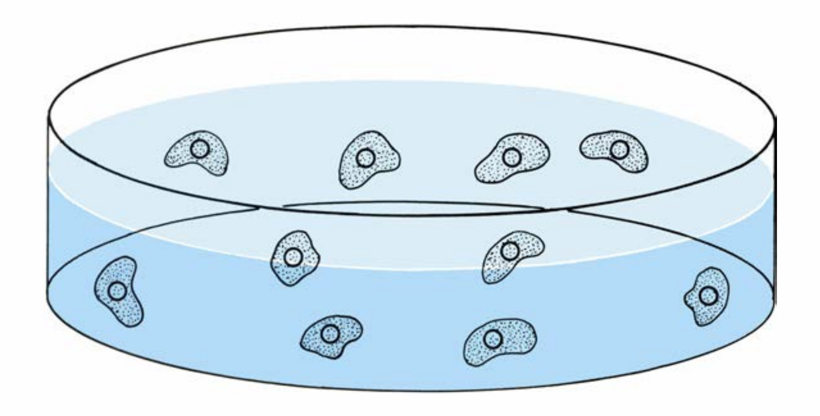
Todo lo cual suscita la cuestión de cómo evolucionan exactamente los glovitos en mi caldo computarizado. Periódicamente, un rayo cósmico atraviesa la sopa e impacta al azar en un cromosoma, en un lugar igualmente aleatorio. El resultado de ello es que un gen específico pasa de un alelo a otro. Por ejemplo, en el siguiente cromosoma, que pertenece a un glovito de cuatro estados, el gen situado en el locus 3 controla el símbolo de salida correspondiente a la transición desde el estado *A*, cuando a la criatura le llega un 1:

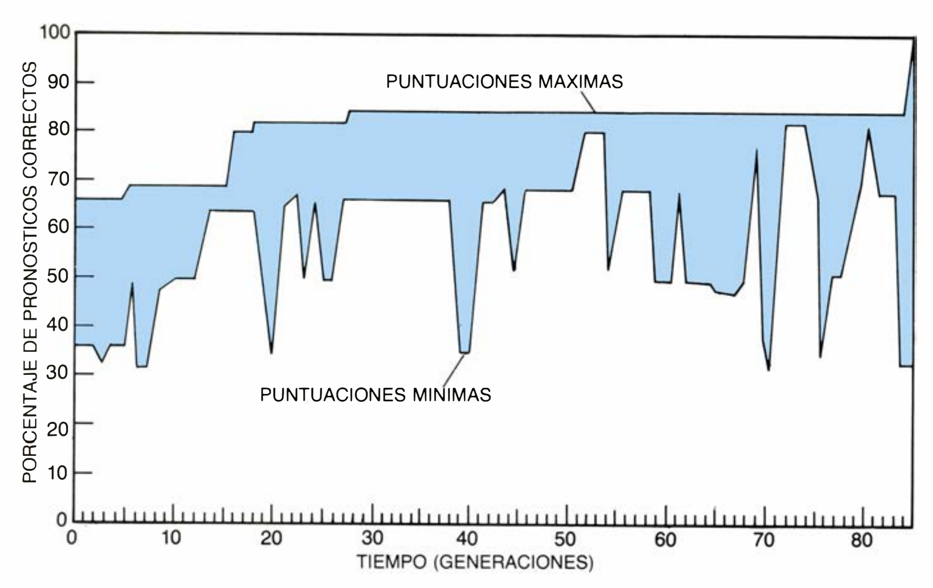
0 D 1 C 0 D 0 B 1 A 0 C 1 B 1 A

Un rayo cósmico impacta en este gen, y modifica ligeramente su cromosoma:

0 D 0 C 0 D 0 B 1 A 0 C 1 B 1 A

La otra fuente de variación del acervo genético de los glovitos es el apareamiento. Durante la estación de procreación, el glovito de máxima puntuación intercambia genes con otro elegido al azar. La descendencia porta un cromosoma compuesto. Parte de él proviene del progenitor superior, la otra, del ganador en la lotería de apareamiento. La composición se asemeja al fenómeno de entrecruzamiento que se da en los cromosomas reales. En los cromosomas de los glovitos, el entrecruzamiento puede ilustrarse combinando el primero de los cromosomas





2. En un sopicaldo con 10 glovitos (**arriba**) *aparece por evolución un pronosticador perfecto* (**abajo**)

antes mencionados (sin alteración) con otro:

1 A 1 B 0 D 1 A 0 C 1 D 1 B 0 C
↑ ↑

Las flechas indican los puntos de entrecruzamiento, elegidos de forma estocástica. El cromosoma descendiente es idéntico al del segundo progenitor hasta llegar al primer punto de cruzamiento. Entre dicho punto y el segundo es idéntico a la correspondiente porción del cromosoma del primero. Tras el segundo punto es nuevamente idéntico al cromosoma del segundo progenitor [*véase la figura 3*]:

1 A 1 C 0 D 0 B 0 C 1 D 1 B 0 C

Antes de decidirme a redactar y poner a prueba el programa primordial era yo un tanto escéptico acerca del valor del apareamiento por cruce. Descubrí con sorpresa, sin embargo, que si el primer progenitor es razonablemente acertado en sus pronósticos, la descendencia tiende a serlo también.

Los lectores pueden juzgar por sí mismos preparando un programa que llamaremos AUTOSOPA. Al listarlo, el programa no excede demasiado de una página. Consta de cuatro módulos inmersos en un bucle. Es preciso establecer un límite que defina la puntuación máxima. En tanto la puntuación máxima sea inferior al límite, el programa debe continuar funcionando a través de los cuatro módulos.

En el primero de los módulos se determina la puntuación de los 10 glovitos sobre una secuencia de 100 símbolos ambientales. El segundo módulo sirve para identificar los glovitos que obtienen las puntuaciones mínima y máxima resultante. En el tercer módulo, el glovito de máxima puntuación es apareado con un compañero seleccionado al azar. El fruto de esta unión remplaza al glovito de puntuación mínima. En el cuarto y último módulo llega un rayo cósmico, que alcanza a un glovito cualquiera y provoca una mutación. Justamente antes de que el programa vaya a invocar al tercer módulo (el de apareamiento) se selecciona un número al azar. Si el número queda por debajo de cierto umbral, se pasa por alto el módulo de apareamiento y se ejecuta inmediatamente el módulo de mutación. Este umbral puede ajustarse al valor que se desee. Empero, ciertos valores dan mejores resultados que otros. Si el módulo de apareamiento es ejecutado con excesiva frecuencia, la pequeña población queda rápidamente dominada por los genes del glovito de máxima puntuación. El fondo de recursos genéticos pierde variedad, y la evolución se decelera hasta un penoso arrastrar, o llega a detenerse completamente. En todo caso, la evolución se va haciendo cada vez más lenta conforme aumentan las puntuaciones. El glovito de máxima puntuación permanece en escena durante períodos cada vez más largos, porque resulta cada vez menos probable que lleguen a evolucionar glovitos superiores a él.

En AUTOSOPA resulta conveniente utilizar cuatro tablas. Se llama *crom, estado, tanteo* y *e*. La tabla *crom* es una matriz bidimensional que consta de 10 glovitos y 16 genes. *Crom* (*i,j*) corresponde al *i*-ésimo gen del *j*-ésimo glovito. Las tablas de estado y tanteo contienen el estado actual y la puntuación de los 10 glovitos. La cuarta tabla, *e*, contiene la cadena de caracteres básica utilizada para generar los símbolos del entorno. Esta lista se da mediante el teclado, al comienzo del programa.

La evaluación de los glovitos se efectúa merced a un bucle doble. El bucle externo engendra 100 símbolos ambientales y el bucle interior incrementa el tanteo de cada glovito si éste atina a pronosticar correctamente el signo siguiente. Podemos someter adecuadamente a prueba los glovitos de cuatro estados sobre entornos de período seis. Ello supone un problema de mediana dificultad. La evolución de predictores perfectos en ambientes de período ocho puede exigir sesiones de funcionamiento del programa de todo un día; en cambio, los entornos de período cuatro apenas si suponen dificultad alguna. En este primer módulo conviene utilizar dos trucos. El primero de ellos permite hacerse con el siguiente símbolo ambiental a partir del índice *i* del bucle exterior, calculando *i* módulo seis, o sea, el resto de la división de *i* entre seis. El número obtenido puede utilizarse como índice en la tabla *e*. Al ir *i* recorriendo los valores de 1 a 100, el índice computado va repasando la ta-

bla e una y otra vez, generando así la secuencia adecuada de símbolos ambientales. Conocido el índice del símbolo actual resulta fácil calcular cuál será el índice siguiente, y consultar la tabla. Tal símbolo se compara entonces, por turno, con la predicción hecha por cada glovito.

El segundo truco permite al programa hallar rápidamente el estado consecutivo del glovito, y determinar cuál será su salida, sin más que inspeccionar el cromosoma. En lugar de representar los cuatro estados por A, B, C y D, en la tabla *estados* se utilizan como elementos los números 0, 1, 2 y 3. Llamando *simb* al símbolo de ambiente, la salida engendrada por el i-ésimo glovito puede hallarse usando primero una sencilla fórmula:

$$l = 4 \times estado(i) + 2 \times simb.$$

Seguidamente es preciso identificar el contenido de $crom(i,l)$. El locus l del cromosoma del i-ésimo glovito emite su salida cuando la criatura se encuentra en el estado i y está recibiendo el símbolo de entrada *simb*. El estado siguiente ocupa el locus $l+1$.

El módulo que determina los glovitos que encabezan y cierran la tabla de clasificación por tanteo se vale de un ejercicio frecuente en los cursos elementales de programación: dada una tabla de n números, escribir un programa que determine el máximo de todos ellos. Para resolverlo se da a una variable llamada *max* el valor inicial 0, y se revisa la tabla mediante un bucle simple. Cada elemento de la tabla se va comparando con *max*. Si el elemento resulta ser mayor que *max*, el valor de ésta es remplazado por el de tal elemento. Convendría también guardar el índice del elemento dentro de la tabla. El mismo programa, con una leve modificación, sirve para hallar el tanteo mínimo. Esta vez será preciso dar a una variable, *min*, el valor inicial 100 e irla remplazando por el valor de los elementos que vayan siendo menores que ella.

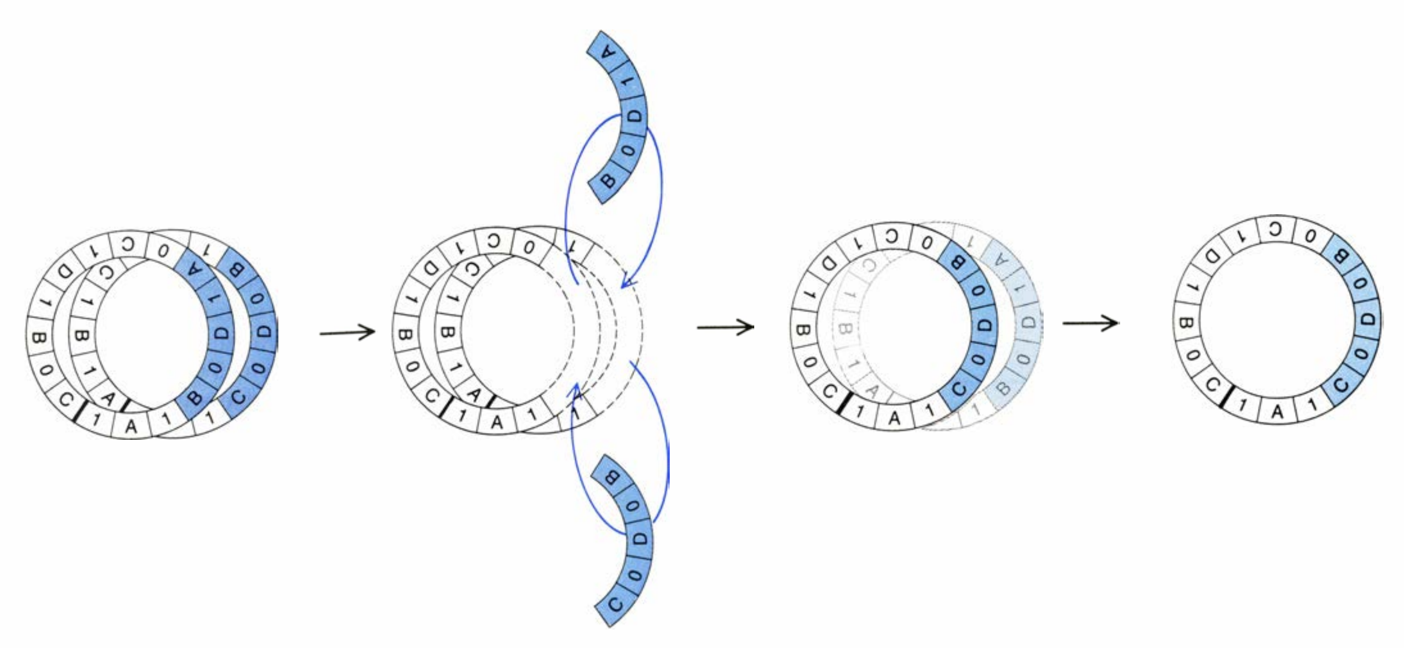
El tercer módulo se encarga de aparear el glovito de máxima puntuación con un individuo seleccionado al azar en la población. La única dificultad que plantea este segmento del programa estriba en la selección de los puntos de cruzamiento. Me parece que lo más sencillo es seleccionar al azar dos enteros c_1 y c_2 en el intervalo de 1 a 16. En el caso de que c_1 fuese mayor que c_2, se intercambiarían sus valores. Los lectores podrán comprobar que con un poco de *finesse* bastan tres bucles, que vayan de 1 a c_1, de c_1 a c_2, y de c_2 a 16, para disponer de toda la maquinaria necesaria para trasladar elementos de *crom* situados en las hileras inseminantes hasta la fila destinataria, que está ocupada por el glovito condenado a desaparecer, por tener puntuación mínima.

En el cuarto módulo tendrían que seleccionarse al azar un índice de glovito y un *locus* dentro de su cromosoma. La paridad de *locus* determina si quien ha de mutar habrá de ser un gen de estado o un gen de salida. Si el valor es 0, será preciso sumar 1 módulo 2 al número ya almacenado allí. Por así decirlo, este proceder "voltea" el bit. Si el valor, módulo 2, de *locus* es 1, es preciso sumar 1 módulo 4 al elemento de la tabla. De este modo se cambia el valor almacenado.

¿He hecho trampa? Sin duda, mal puede decirse que sea aleatorio el sistemático cambio de estado, de 0 a 1, y luego a 2 y a 3, y nuevamente a 0. Yo replico que es suficientemente aleatorio: el número de estados es lo suficientemente pequeño para que uno no pueda esperar que el resultado final del programa sea muy diferente del obtenido cuando prevalezcan estados seleccionados más aleatoriamente. En realidad, yo había hecho también un poco de trampa al elegir c_1 y c_2 tan descuidadamente: tal método garantiza a ciertas subcadenas ventajas con respecto a otras. Pero, lo mismo que antes, me parece que las diferencias entre AUTOSOPA y un procedimiento de selección del cruzamiento estadísticamente corregido serían muy ligeras. Tanto de un modo como de otro se hacen con los genes tantos malabarismos y se han fragmentado tantas veces los cromosomas, que al glovito de cabeza de tabla le costaría mucho trabajo reconocer a sus propios nietos.

Las únicas partes de AUTOSOPA todavía no especificadas son su comienzo y su final. Los glovitos que inicialmente ocupen el caldo primordial deberían haber sido seleccionados al azar. Para determinar cada uno de los genes de cada glovito se tendría que elegir un entero dentro de la gama apropiada, y serle asignado a ese gen. Finalmente, cuando un glovito exceda por primera vez el límite establecido en el bucle exterior, AUTOSOPA debería imprimirlo.

Es preciso prevenir a los lectores dis-



3. El cruzamiento de dos cromosomas y el cromosoma de la progenie resultante (**derecha**)

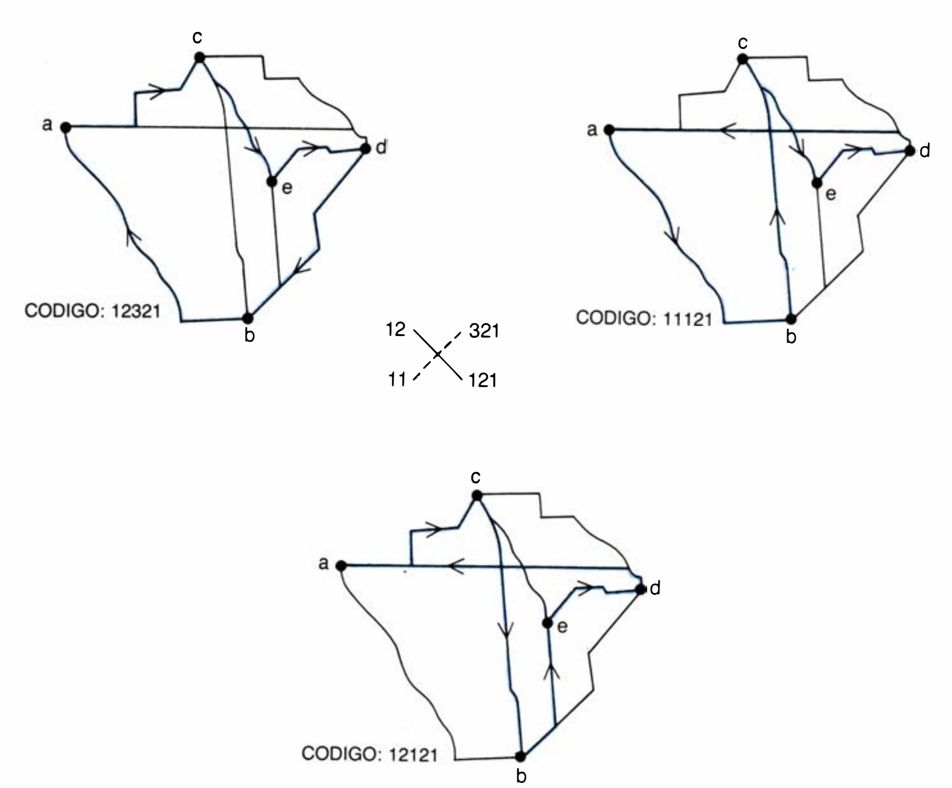
puestos a embarcare en esta aventura genética de que hay en ella no poca exploración. Puede ocurrir que algunos de los exploradores lleguen a padecer de adicción. La presencia de evolución, y su ritmo, son cuestiones a elucidar. Cuando el período del entorno sea demasiado largo para que llegue a evolucionar un predictor perfecto de cuatro estados, ¿hasta qué punto llegan a aproximársele los glovitos? ¿De qué forma influyen los cambios de la longitud del período en la duración del tiempo requerido para que llegue a evolucionar un predictor perfecto? No hay nada en la descripción de AUTOSOPA que impida generalizar el programa para glovitos de 5 ó 6 estados. Se puede incluso modificar el programa para explorar ambientes no periódicos, o bien ambientes que ocasionalmente cambien la serie básica de símbolos.

El caldo de autómatas se inspiró en un libro aparecido a comienzos del decenio de 1960. Titulada *Artificial Intelligence through Simulated Evolution*, la obra describe una serie de experimentos sobre la evolución de los autómatas, efectuados por Lawrence J. Fogel, Alvin J. Owens y Michael J. Walsh. A los autómatas se les pedía que pronosticasen secuencias periódicas, y se les permitía evolucionar de modo similar al de nuestros glovitos. Sin embargo, en aquel austero estudio, a los autómatas no se les permitía el apareamiento ni el entrecruzamiento de genes.

Fue Holland quien me sugirió que añadiera a la sopa de autómatas el ingrediente del entrecruzamiento genético. Como ya hice notar, Holland es el padre reconocido del algoritmo genético. Los especialistas de este sistema, cuyo número crece constantemente, se reunieron en una primera conferencia a gran escala, debidamente financiada, en la Carnegie-Mellon University. Allí analizaron un amplio abanico de teoría y aplicaciones. Un problema, analizado en diversos artículos, puede servir de interesante introducción a la materia de la programación genética.

Se llama "problema del viajante", y lanza el siguiente reto: dado el mapa de *n* ciudades, enlazadas por una red de carreteras, hallar el mínimo circuito que pase por todas ellas. Tal circuito le serviría al viajante para reducir al mínimo sus gastos de desplazamiento.

Resulta del todo posible desarrollar por evolución un recorrido de longitud mínima, exactamente de igual modo que llegaron a evolucionar, hasta convertirse en predictores perfectos, glovitos que distaban de serlo. Cada circuito podría codificarse en un cromosoma. Los más cortos podrían entonces aparearse con otros, con la esperanza de lograr descendientes más breves todavía. El entrecruzamiento produce los cromosomas de la progenie.



*4. Dos vendedores parentales (**arriba**) y un descendiente (**abajo**) obtenido por encruzamiento genético*

La tarea de elegir una buena representación para los circuitos es tan seductora como engañosa. Por ejemplo, si utilizamos una simple lista de las ciudades, en el orden en que se han visitado, la descendencia puede que ni siquiera sea un circuito. Para eludir esta dificultad, los autores de un artículo, John Grefenstette, Rajeev Gopal, Brian Rosmaita y Dirk Van Gucht, de la Universidad de Vanderbilt, han propuesto un ingenioso cromosoma. La representación de una gira a través de cinco ciudades, así *a, c, e, d, b*, resulta ser 12321. Para obtener tal serie numérica se recurre a un cierto orden convencional para las ciudades, el alfabético *a, b, c, d, e*, pongamos por caso. Dado un circuito como *a, c, e, d, b*, se van retirando sistemáticamente las ciudades de la lista patrón, en el orden en que se presentan en el circuito. Se retira *a*, después *c, e*, y así sucesivamente. Conforme va retirándose cada ciudad de la lista convencional, se anota su posición en lo que quede de lista justamente antes de extraer la letra: la *a* es la primera, la *c*, segunda, la *e*, tercera, la *d*, segunda y, finalmente, la *b* es primera. Emerge así el cromosoma 12321. Lo interesante es que por entrecruzamiento de dos de esos cromosomas se obtiene siempre un circuito [*véase la figura 4*]. Con esa representación pueden, por así decirlo, "criarse" ahora nuevos recorridos, buscando los más adecuados.

Casi todos los especialistas del arte de los algoritmos genéticos están dispuestos a conceder que el problema del viajante es uno de los más grandes retos que tienen planteados. Si bien los experimentos efectuados con la representación recién descrita no han dado resultados demasiado estimulantes, existen otros algoritmos genéticos que aportan un mejor rendimiento en este problema.

Empero, ningún algoritmo genético ha sido capaz hasta ahora de conquistar una solución al problema del viajante, en ninguno de los sentidos de "solución" aceptados comúnmente. Ello es debido, sin duda alguna, a la intrínseca dificultad del problema. Por tratarse de uno de los problemas que en teoría de computación se denominan NP-completos, es muy posible que sea su destino el permanecer por siempre insoluble en la práctica.

Desde el pasado mes de octubre, en que fue presentado en esta sección el programa MANDELZOOM, ha tomado cuerpo en centenares de casas, escuelas y oficinas de trabajo. Aunque

el programa parece haber infundido a los adultos un cierto respetuoso temor, intrigado a los adolescentes y asustado a unos pocos niños pequeños, para sorpresa mía el correo referente a los diagramas de iteración, que era tema secundario en mi artículo, ha sido casi tan abundante como el alusivo al conjunto de Mandelbrot.

Muchos han sido los lectores que se han perdido entre las coloreadas fragosidades del conjunto de Mandelbrot, escudriñándolo cada vez más profundamente y con mayores ampliaciones. Algunos, careciendo de equipo para crearlas por sí mismos, pero decididos a poseer sus propias imágenes en color, se las han encargado a John H. Hubbard, de la Universidad de Cornell. Según me dice Heinz-Otto Peitgen, el explorador de conjuntos de Mandelbrot cuyas imágenes embellecieron las páginas del número de octubre, también aquéllas están a la venta.

Debí haber pensado en la posibilidad de utilizar diversos matices de gris, en atención a los lectores cuyo equipo estuviera limitado a blanco y negro. Las imágenes que se obtienen pueden ser casi igual de sugestivas que las correspondientes en color. Las mejores imágenes en gris son obra de David W. Brooks, quien para computar y representar sus imágenes trabaja con el equipo instalado en Prime Computer, Inc., de Framington, Massachusetts. En los fabulosos y delicados torbellinos de entreluces, los distintos matices de gris se consiguen mediante diminutos cuadraditos negros de cierto tamaño; se trazan éstos mediante una impresora a láser. Brooks ha tratado de encontrar los diminutos filamentos que, según se cree, conectan los Mandelbrot en miniatura con el conjunto principal. Hasta ahora no han aparecido en ninguna de las ampliaciones utilizadas por Brooks. Mandelbrot le ha prevenido de que posiblemente sean infinitesimales.

Quienes no dispongan de equipos tan perfeccionados pueden trabajar todavía con matices de gris en un monitor blanco y negro. John B. Halleck, de Salt Lake City, hace variar la densidad de puntos por píxel para expresar los diferentes matices.

Un enfoque distinto consiste en trazar contornos en blanco y negro. Yekta Gursel, de Cambridge, Massachusetts, ha generado representaciones del conjunto de Mandelbrot capaces de competir con las de Brooks. Gursel sustituye un espectro discreto de colores por bandas alternantes de blanco y negro. Gary J. Shannon, de Grants Pass, Oregon, propuso la misma técnica y Victor Andersen, de Santa Clara, California, la ha llevado al extremo. Andersen propone cambiar de blanco a negro, o viceversa, cada vez que la variable *recuento* cambie al pasar de un píxel a su vecino.

Vale la pena citar otras dos exploraciones. James A. Thigpenn IV, de Pearland, Texas, en lugar de colores utiliza alturas. El conjunto de Mandelbrot se convierte en una gran meseta vista desde un ángulo, con un complicado sistema de agudas colinas que se aproximan a la meseta en diversos lugares. Richard J. Palmaccio, de Fort Lauderdale, prescinde además del conjunto. El interés de Palmaccio está en ir siguiendo la trayectoria de los números complejos individuales a lo largo de las iteraciones. En la vecindad de la frontera, los movimientos coreográficos de éstos crean pasos de ballet, en espirales, o trazan círculos, como en la sardana.

La función $z^2 + c$ genera el conjunto de Mandelbrot. Como es natural, se pueden utilizar otras funciones, pero producen otros conjuntos. Por ejemplo, Bruce Ikenaga, de la Universidad Case Western Reserve, ha explorado lo que parece ser un cactus cúbico. La función $z^3 + (c - 1)\,z - c$ produce un conjunto espinoso, rodeado de misteriosas galaxias espirales en miniatura, cuyo aspecto causa una cierta desazón (al menos en blanco y negro puros).

Tampoco faltan los misterios en los diagramas de iteración: al elevar al cuadrado los enteros módulo n, cada número emigra, en efecto, y pasa a ocupar la posición de otro. El diagrama de iteración aparece cuando cada número se remplaza por un punto y, cada movimiento migratorio, por una flecha. Planteaba yo varias preguntas relativas a tales diagramas. ¿Cuántas componentes tienen? Los lectores me enviaron diagramas acreditativos de sus exploraciones para diversos valores de n.

Los diagramas más grandes los preparó Rosalind B. Marimont, de Silver Spring, Maryland, quien examinó los enteros módulo 1000, e informó de la existencia de cuatro pares de componentes en el diagrama de iteración resultante. Cada componente exhibía un único atractor, como de costumbre; los atractores más grandes constaban de 20 números. Siendo matemática, podemos permitir a Marimont conjeturar que los enteros módulo 10^k producirán $k + 1$ pares de componentes, y que los atractores máximos constarán de $4 \times \times 5^{k-2}$ números.

Stephen Eberhart, de Reseda, California, investigó el caso de que n fuera un primo de Fermat (un número primo de la forma $2^{2^k} + 1$). Aquí, el número 0 forma por sí solo un atractor; todos los números restantes se encuentran en un único y frondoso árbol. Un amigo suyo, especialista en teoría de números, afirma que siempre se dará este caso para los primos de Fermat, y que el árbol es binario: a cada punto de su interior llegan siempre dos flechas.

Lo mismo que los números, los diagramas de iteración pueden multiplicarse. Si n es producto de dos números primos entre sí, que llamaremos p y q, el diagrama de iteración para los enteros módulo n es el producto, respectivamente, de los diagramas correspondientes a p y a q. Esta interesante observación se debe a Stephen C. Locke, de la Universidad Florida Atlantic. Locke ha descrito también una fascinante relación entre el n-ésimo diagrama de iteración y un diagrama de tipo aparentemente distinto, en el cual los números, en lugar de elevarse al cuadrado, meramente se duplican. Cuando n es primo, el diagrama de este último, correspondiente a los enteros módulo $n - 1$, es el mismo que nuestro n-ésimo diagrama de iteración, a excepción de un único número aislado, que forma un atractor por sí solo. Noam Elkies, de Nueva York, hizo en buena medida esta misma observación, si bien sus consideraciones se fundan en la teoría de números.

Frank Palmer, de Chicago, puso a punto una poderosa herramienta para analizar los diagramas de iteración (cuadráticos). Al parecer, todos los árboles ligados a un atractor dado son isomorfos, lo que significa que tienen precisamente la misma forma.

Finalmente, Bruce R. Gilson, de Silver Spring, Maryland, y Molly W. Williams, de Kalamazoo, Michigan, estudiaron una generalización totalmente diferente de los números del 0 al 99. Pueden éstos considerarse números en bases distintas. Por ejemplo, tomándolos como números en base 3, tendríamos que contar 00, 01, 02, 10, 11, 12, 20, 21, 22 antes de volver a llegar a 00. Tales números producen también diagramas de iteración similares a los generados por los enteros módulo n. Gilson demostró que los diagramas tienen siempre componentes emparejados cuando n es par, pero no múltiplo de 4.

Se cometió un error en el diagrama de iteración presentado en el número de octubre para los números de 0 a 99. Faltaban dos flechas en las dos componentes y la dirección de uno de los atractores era contraria a la debida.

Taller y laboratorio

Cómo observar el Halley durante los próximos meses

Jearl Walker

Actualmente el cometa Halley se encuentra viajando por la región de los planetas interiores y circunvalando el Sol. Recibió su nombre de Edmund Halley, astrónomo inglés del siglo XVII, el primero en descubrir que visita periódicamente nuestro sistema solar. Valiéndose de las leyes de la mecánica y de procedimientos esbozados con anterioridad por su coetáneo Isaac Newton, pudo averiguar que los cometas observados en 1531, 1607 y 1682 presentaban órbitas similares. Sospechando que todos esos registros correspondieran a un mismo objeto, predijo que el cometa volvería en 1758. Y así fue, tal como observó un astrónomo aficionado en la noche de Navidad. De hecho, investigaciones recientes revelan que el cometa Halley fue visto ya en el año 240 a. C. y que, desde entonces, se ha venido observando cada 76 años aproximadamente. Este regreso es el cuarto desde que Halley propuso su teoría.

El cometa se dejará ver, si bien tenuemente, por los habitantes del hemisferio boreal. Insistiré, pues, en las observaciones que serán posibles en esas latitudes medias. Mi información procede del Comité Internacional de Observación del Halley, del Laboratorio de Propulsión a Chorro del Instituto de Tecnología de California y de libros de Robert D. Chapman y John C. Brandt, del Centro de Vuelos Espaciales Goddard de la Administración Nacional de la Aeronáutica y del Espacio (NASA), así como de investigaciones de David W. Hughes, de la Universidad de Sheffield.

Dada la palidez del cometa, sólo podrá verse en caso de que el fondo del firmamento nocturno sea oscuro. Por culpa del fulgor de la luz artificial dispersada por el aire, no es probable que se aprecie desde zonas residenciales; tampoco en noches de luna brillante. Un buen indicio de que un lugar es apto para la observación es que desde éste se vea la Vía Láctea. Se dejará entonces que los ojos se acostumbren a la oscuridad durante 15 minutos al menos.

Aunque podría distinguirse a simple vista, unos binoculares mejorarán notablemente la observación ya que captan más luz. Servirán en especial los de siete aumentos con un objetivo de 50 milímetros de diámetro, aunque también puede recurrirse a otros de 35 milímetros. Unas veces, el cometa ocupará un ángulo tan grande en el campo de visión que, para verlo todo, quizá convenga efectuar un barrido con los binoculares. No es adecuado el telescopio de muchos aumentos, pues la parte del cometa que permite ver es demasiado pequeña para que contraste con suficiente intensidad sobre la oscuridad de fondo. Puede probarse con un telescopio gran angular a su aumento mínimo.

En las figuras 3 y 4 se ilustra dónde se encontraba la cabeza del cometa durante diversas noches del mes de diciembre y dónde se hallará en este mes de enero y siguientes para un observador situado a 40 grados de latitud norte (que es, aproximadamente, la latitud de Madrid, Valencia y Barcelona). Para observadores en latitudes mayores, el cometa aparece más bajo sobre el firmamento y sólo cae encima del horizonte durante períodos más cortos. Cada ilustración es una gráfica de la altitud y el azimut del cometa en grados. La altitud se mide a partir del horizonte y el azimut en sentido horario desde el rumbo norte.

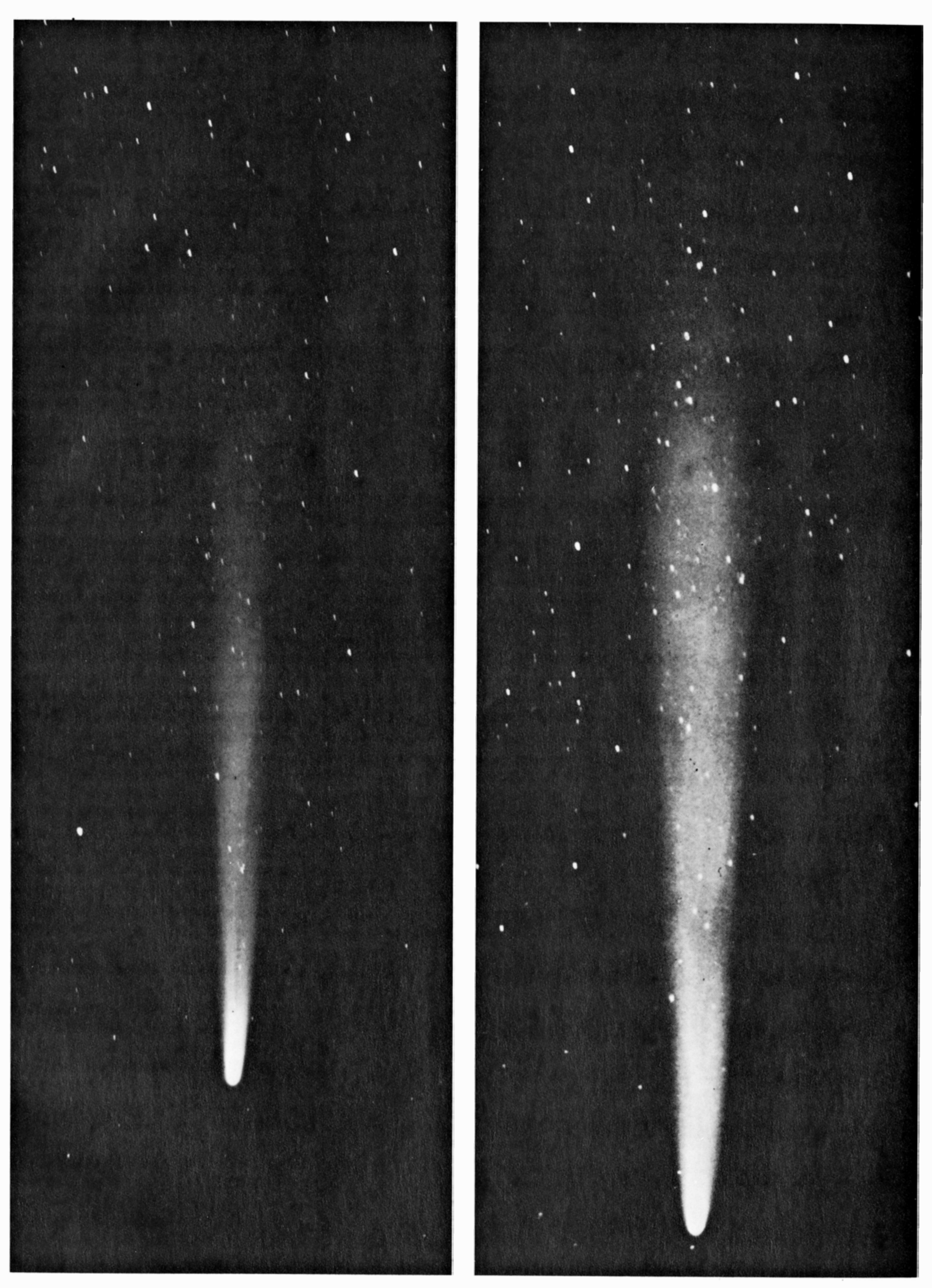
En las gráficas los circulitos señalan la posición de la cabeza del cometa de hora en hora para diversas noches. Las posiciones anteriores a la medianoche están enlazadas con un trazo discontinuo y las posteriores a la medianoche con un trazo continuo. El primer círculo de cada curva muestra la fecha y la hora (en tiempo oficial local en reloj de 24 horas) en que la cabeza del cometa ocupa esa posición. Por ejemplo, la línea más alta de la ilustración superior indica que la cabeza del cometa estuvo a una altitud de unos 38 grados y a un azimut de unos 108 grados a las 17,00 (5,00 de la tarde) del primero de diciembre. Más tarde, esa misma noche, la cabeza alcanzó una altitud máxima de unos 60 grados cuando avanzaba con rumbo sur. Bajó luego hasta ponerse en el firmamento occidental a las 3,00 de la madrugada aproximadamente.

Conforme transcurren las noches de diciembre y enero decrece la altitud de la cabeza del cometa en su aparición nocturna. Hacia el 24 de enero el cometa será visible sólo durante un breve tiempo antes de ponerse, al comienzo del atardecer. Durante la segunda mitad, de diciembre y de enero, la Luna brilla mientras el cometa pasa encima del horizonte, por lo que su contemplación resulta deficiente o imposible. Asimismo, la Luna dificultará la observación hacia finales de febrero, durante la última semana de marzo y en los primeros y últimos días de abril.

A lo largo de febrero y marzo, los períodos de observación serán cortos. Como el cometa se presentará bajo en el cielo, debe buscarse un punto de observación que ofrezca un horizonte bajo y despejado. Durante esos meses, además, apenas resplandecerá, ya que, a causa de su baja posición, su luz atravesará una gran porción de atmósfera terrestre antes de llegar al observador. En algunos lugares, la dispersión de la luz por el aire atmosférico y los aerosoles, durante el largo trayecto, pueden empalidecerlo tanto que lo hagan inobservable. Muchos expertos opinan que la mejor oportunidad para verlo se dará el 15 de abril aproximadamente, cuando su cabeza brille y se presente un poco por encima del horizonte con un firmamento oscuro.

La posición de los cometas y de otros objetos celestes acostumbra a especificarse en ángulos medidos con respecto al plano ecuatorial, un ente imaginario de referencia que se extiende indefinidamente desde el ecuador terrestre hacia afuera. Este plano corta una esfera celeste imaginaria, que es el ente de referencia para ubicar los cuerpos celestes. El plano ecuatorial separa los hemisferios norte y sur de la esfera celeste. El Sol cruza el plano ecuatorial dos veces al año, en los equinoccios vernal y otoñal.

En la figura 5 se muestra un observador situado en el centro del plano ecuatorial, el cual está inclinado con respecto a un plano horizontal que abarca todo su horizonte. (El ángulo de inclinación depende de la latitud del observador: cuanto mayor es ésta, menor es el ángulo.) El horizonte está se-



1. El cometa Halley el 12 y 15 de mayo de 1910, fotografiado desde Honolulú

ñalado con los símbolos N, S, E y O. El campo estelar visto por el observador yace en la esfera celeste, la cual aparenta rotar en torno al polo norte celeste en virtud del giro de la Tierra.

Para ubicar un objeto en la esfera celeste nos servimos de dos coordenadas, la declinación y la ascensión recta. La declinación define el ángulo entre el objeto y el punto más cercano del plano ecuatorial. Se llama ascensión recta al ángulo entre ese punto del plano y la posición del equinoccio vernal en la esfera celeste. Por convenio, el ángulo de ascensión recta se mide en unidades de tiempo: una hora equivale a 15 grados. La utilidad de este sistema de coordenadas reside en que gira en torno al polo norte celeste junto con toda la esfera celeste y, por ello, las coordenadas de cualquier estrella permanecen virtualmente constantes toda la noche. Además, estas son independientes de la posición geográfica del observador.

La figura 6 ilustra el movimiento del cometa Halley en la esfera celeste durante estos meses. Hasta el 23 de diciembre la declinación fue positiva: el cometa cayó "por encima" del plano ecuatorial. La declinación se hace negativa cuando el cometa desciende "por debajo" del plano ecuatorial.

El movimiento puede también describirse en función de tres ángulos medidos con respecto al plano de la eclíptica, que es el plano de la órbita que describe la Tierra en torno al Sol [*véase la figura 8*]. La inclinación orbital del cometa es el ángulo que habría que girar el plano de la eclíptica para que coincidiera con el plano orbital del cometa, de suerte que Tierra y cometa orbitaran en torno al Sol en el mismo sentido. El cometa Halley presenta una inclinación orbital de 162 grados.

2. Cabeza del cometa el 8 de mayo de 1910

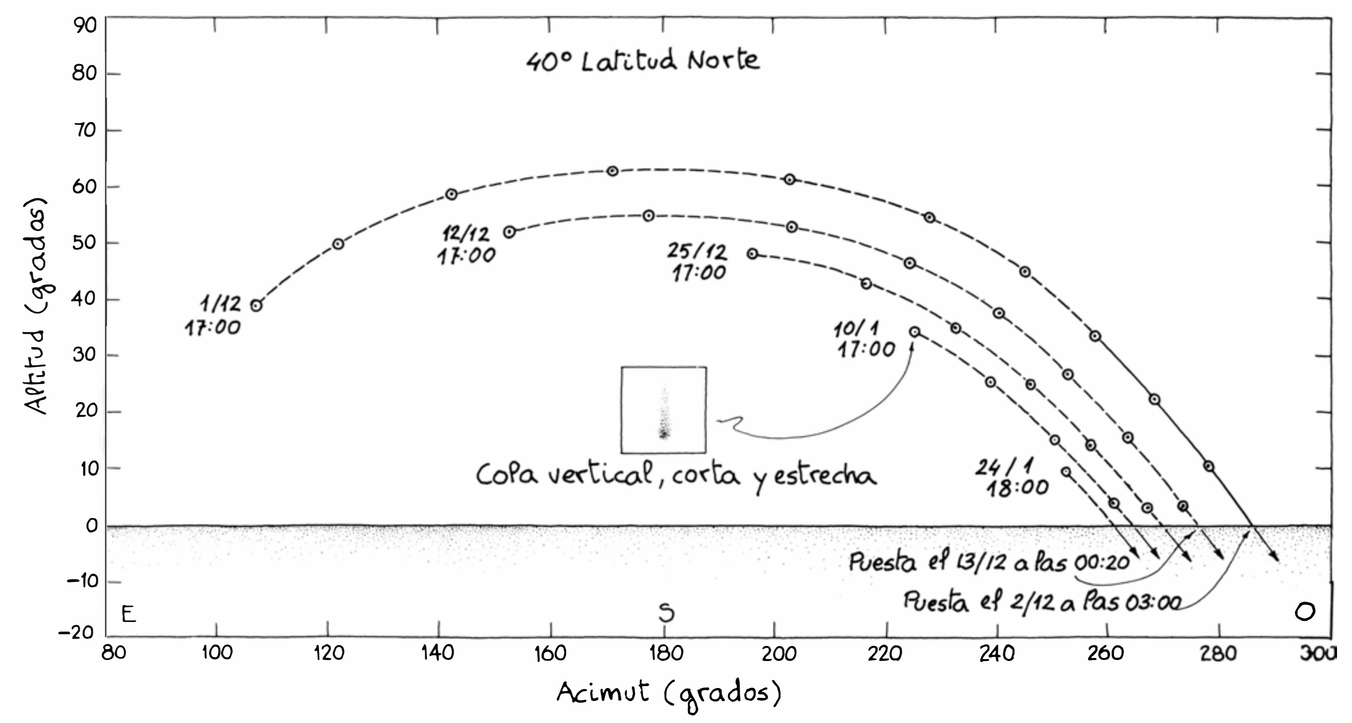
El segundo ángulo de referencia es la longitud del nodo ascendente, el punto por donde el cometa cruza hacia arriba el plano de la eclíptica en su acercamiento al Sol. Este ángulo, de 58 grados para Halley, se mide entre la dirección del equinoccio vernal y una línea que une el Sol con el nodo.

El tercer ángulo está relacionado con el perihelio del cometa (su punto de mayor acercamiento al Sol, que será el 9 de febrero de 1986) y se llama argumento del perihelio. Es el ángulo que subtienden dos líneas que salen del Sol: una que pasa por el perihelio y la otra por el nodo ascendente. Para el Halley este ángulo es de unos 112 grados.

En la figura se muestran además las posiciones aproximadas de la Tierra en los momentos de mayor cercanía al cometa. Antes del perihelio, el acercamiento máximo dióse el 27 de noviembre de 1985, cuando la separación era de unos 93 millones de kilómetros. Después del perihelio, el mayor acercamiento será de unos 48 millones de kilómetros y se producirá el 11 de abril de 1986. (En su aparición de 1910, la separación del cometa fue sólo de 24 millones de kilómetros, por lo que se presentó mucho más luminoso que en la presente visita.) Cuando el cometa alcanza su perihelio resulta inobservable, puesto que su visual y la del Sol difieren en poco, y el resplandor solar nubla la vista de los observadores.

En la figura 9 se representa la órbita completa del cometa Halley en el sistema solar, con su posición en diversos años. Una órbita entera dura 76 años, lo que lo convierte en un cometa de período corto. Las órbitas de los cometas de período largo duran 200 años o más. Ambos tipos difieren en otros aspectos. Los planos orbitales de los cometas de período corto yacen generalmente cerca del plano de la eclíptica y además estos cometas suelen orbitar en torno al Sol en el mismo sentido que los planetas (en sentido antihorario vistos desde encima del plano de la eclíptica). Las órbitas de los cometas de período largo pueden presentar cualquier inclinación y muchas de ellas son recorridas en sentido horario.

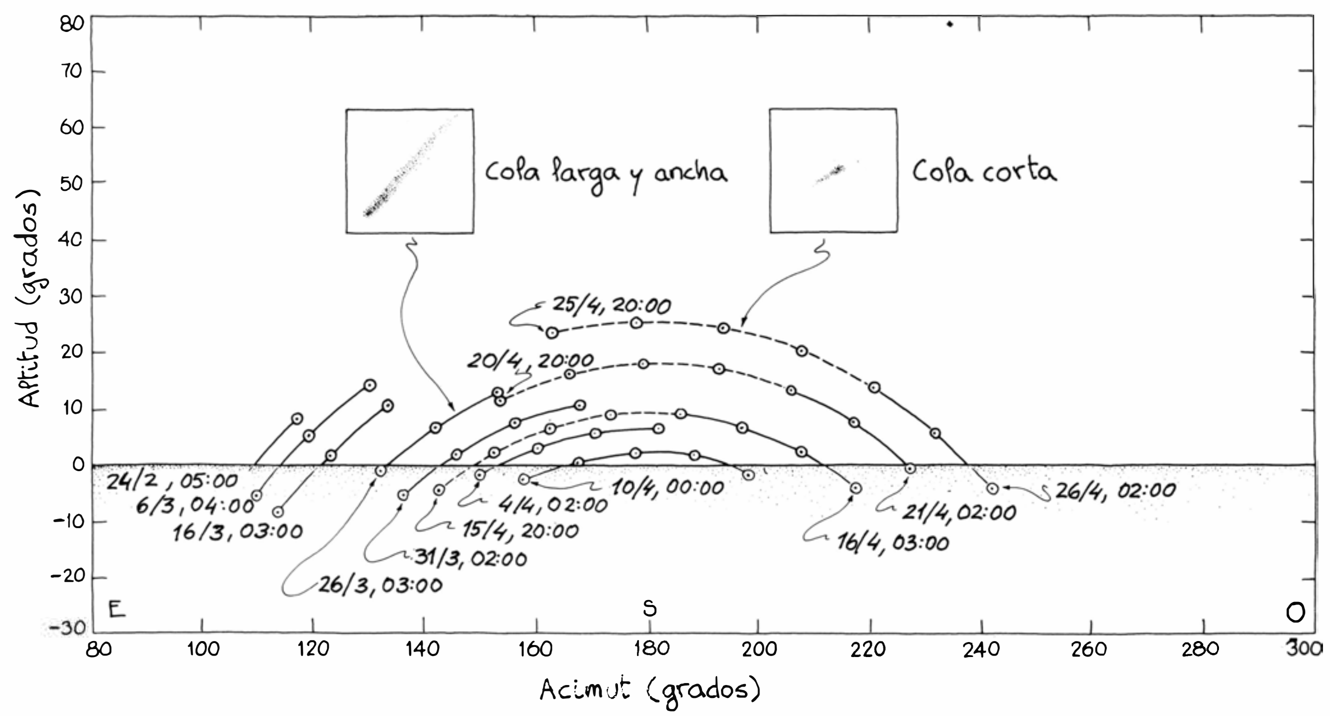
Los cometas proceden, se cree, de una gigantesca nube de objetos que orbitan alrededor del Sol a una distancia de 20.000 a 100.000 unidades astronómicas. (Se llama unidad astronómica, cifrada en 150 millones de kilómetros, a la distancia media entre la Tierra y el Sol.) Esa nube recibe el nombre de nube de Oort, en honor del astrónomo holandés Jan Oort, quien en 1950 señaló su existencia.



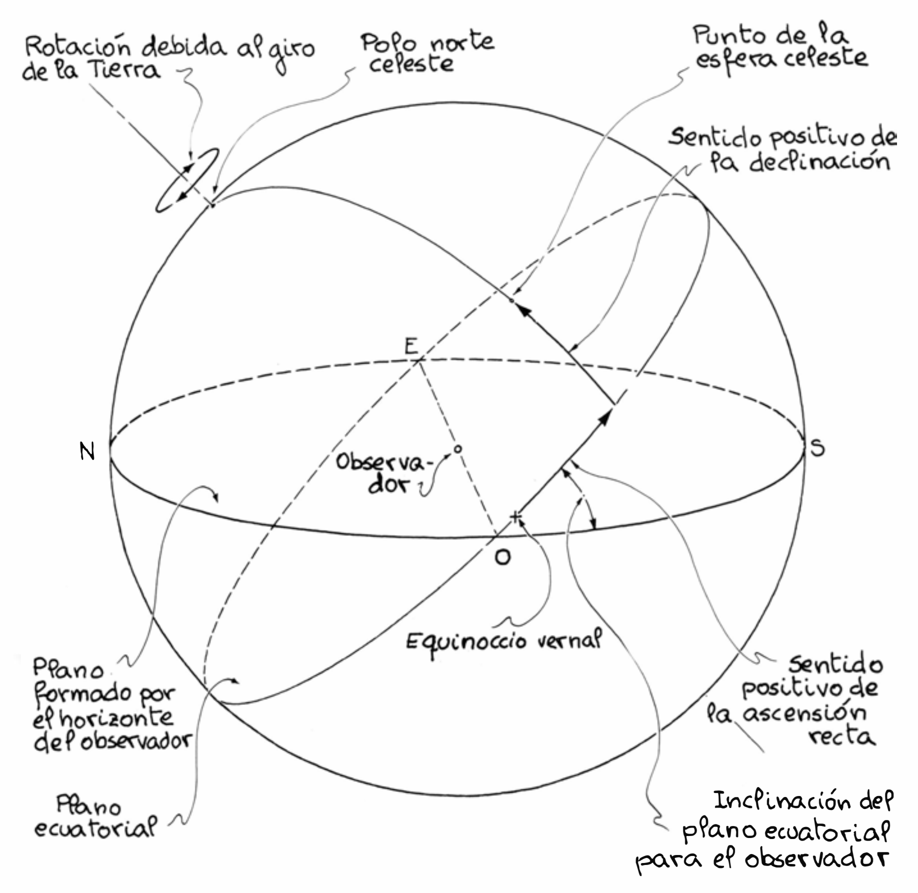
3. Posiciones del cometa en el hemisferio norte durante diciembre de 1985 y enero de 1986

A veces, la atracción gravitatoria de una estrella que pase junto a la nube de Oort hace salir de ella algunos cometas. De éstos, unos se escapan y otros penetran en el sistema solar. De estos últimos, muchos acaban describiendo órbitas grandes de períodos largos. Pero algunos pasan una y otra vez cerca de Júpiter, u otro planeta gigante, por cuya razón la atracción gravitatoria del planeta contrae la órbita hasta que el objeto se convierte en un cometa de período corto. Parece que el cometa Halley pertenece a esta clase.

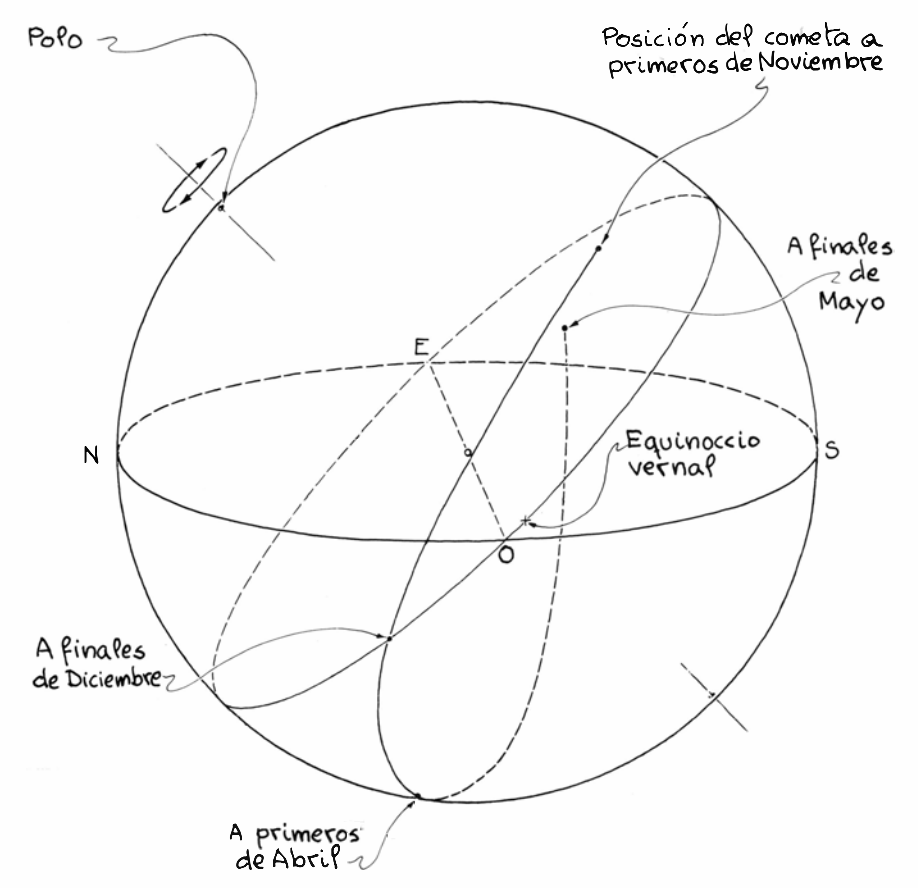
El núcleo de un cometa es una estructura compacta cuya anchura varía desde un kilómetro hasta varias decenas de kilómetros. La forma de un cometa no parece ser esférica, sino más bien irregular, pero nadie ha podido contemplar con nitidez su núcleo dada la pequeñez del objeto. Pese a ello, es posible deducir muchas de sus características. Alrededor del 25 por ciento de la masa la forman partículas de polvo de composición similar a los meteoritos de condritas carbonáceas. Esas partículas tienen un tamaño de 0,1 a 10 micrometros. Se presentan asimismo granos y partículas de polvo mayores. El otro 75 por ciento de la masa se compone de hielo, principalmente hielo de agua y nieve. Hay cometas que



4. Dónde estará el cometa en febrero, marzo y abril de 1986



5. Sistema de coordenadas basado en el plano ecuatorial



6. Movimiento del cometa Halley en la esfera celeste

pueden también albergar un núcleo rocoso. En la disgregada envoltura de cristales de hielo y nieve hay bolsas que contienen diversas moléculas: amoníaco, metano y anhídrido carbónico.

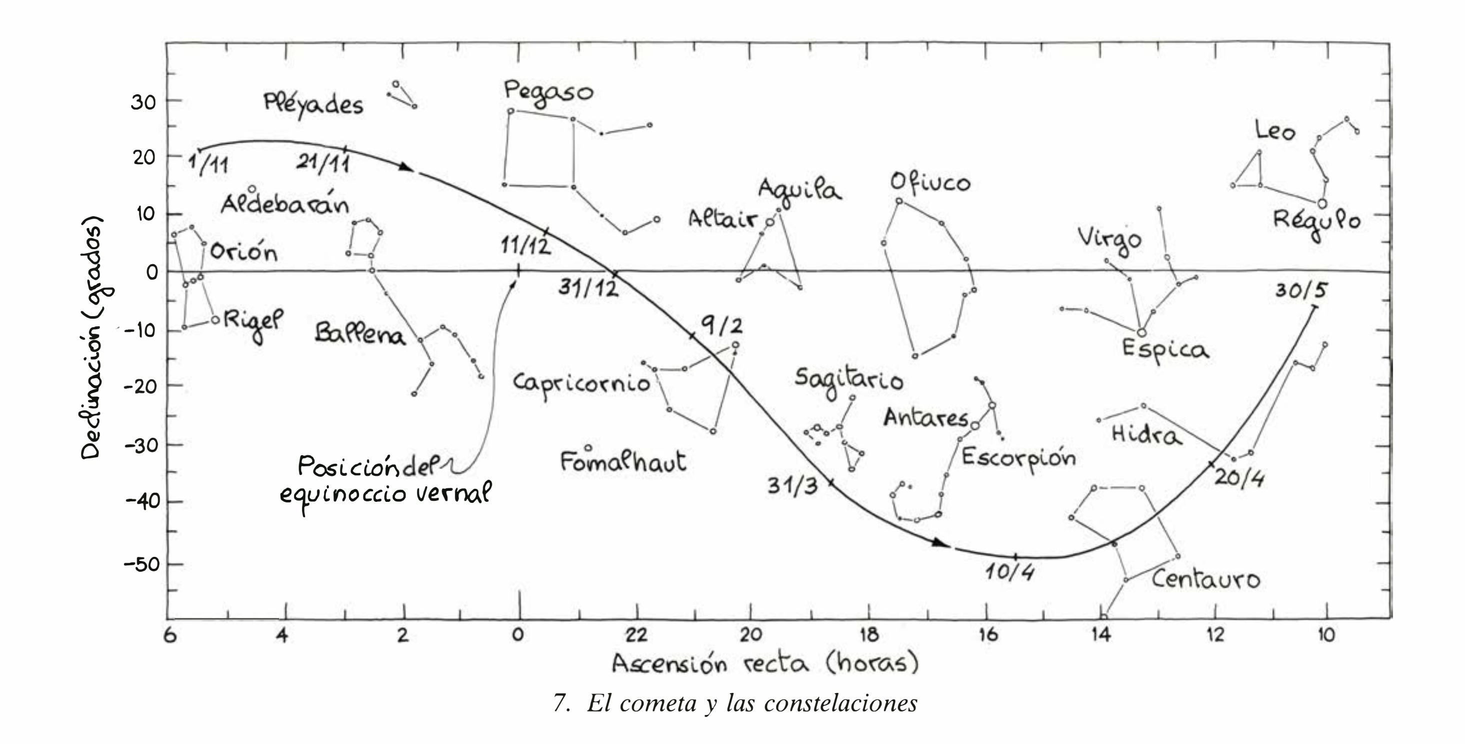
Cuando un cometa se acerca al Sol, el calor en la superficie enfrentada a éste transforma el hielo directamente en gas (proceso llamado sublimación). Así escapan las moléculas atrapadas; es entonces cuando la luz ultravioleta procedente del Sol las ioniza para formar moléculas "hijas" más sencillas, átomos e iones. El examen espectroscópico de los cometas revela la presencia de moléculas de cianógeno (CN) en cometas a una distancia del Sol de hasta tres unidades astronómicas.

Conforme el objeto se acerca al Sol, los registros espectroscópicos exhiben más líneas de emisión, lo que indica que el número de molécuas hijas crece rápidamente. Estas moléculas absorben energía de ciertas longitudes de onda y, luego, la emiten en forma de luz, bien a la misma longitud de onda (proceso conocido como fluorescencia resonante) o a una longitud de onda más larga (fluorescencia). Por otro lado, las moléculas de agua liberadas se disocian para formar una enorme nube de hidrógeno (H) y otra más pequeña de hidroxilo (OH) alrededor del núcleo del cometa. El diámetro de la nube de hidrógeno puede llegar a medir 0,1 unidades astronómicas.

La superficie externa del núcleo, cuando se calienta y pierde hielo, se transforma en una capa esponjosa de polvo cuyo espesor es de uno a 10 centímetros. Si el núcleo gira en torno a sí mismo de modo que toda la superficie se exponga a la luz solar, la capa esponjosa cubrirá todo el núcleo. Esa capa aísla el hielo más profundo y es probable que el centro del núcleo permanezca a temperaturas cercanas a −150 grados Celsius.

Las partículas de polvo de la superficie se desgajan gradualmente y son expelidas por la presión de las moléculas gaseosas que libera la sublimación del hielo en la frontera entre la capa de polvo esponjosa y el hielo en polvo. A la vez que se erosiona la superficie, la frontera se desplaza hacia adentro, manteniéndose el espesor de la capa.

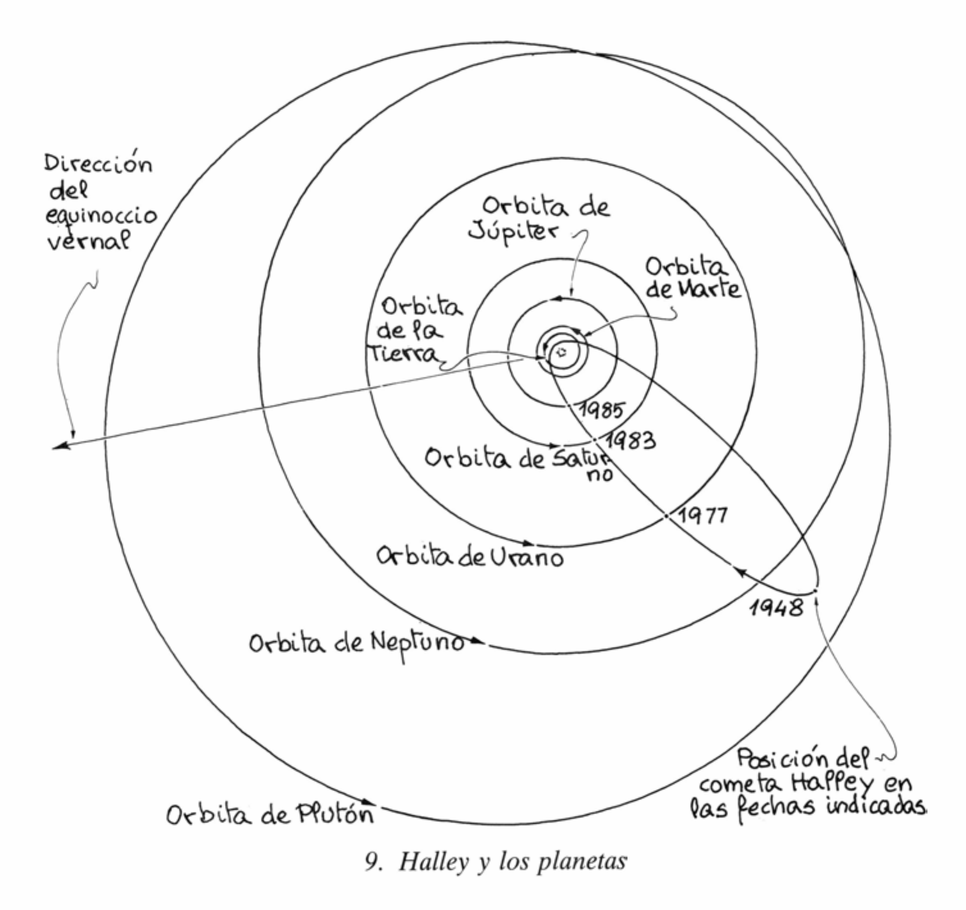
El polvo y el gas neutro liberados por el núcleo forman una nube elíptica y resplandeciente, llamada coma, cuyo diámetro puede llegar a los 100.000 kilómetros. La coma aumenta de diámetro hasta que el cometa se encuentra a 1,5 o 2 unidades astronómicas del Sol. A partir de ahí, la coma mengua



7. El cometa y las constelaciones

Cola de plasma
Cola de polvo
Orbita de la Tierra
Máxima proximidad el 11 de Abril de 1986
En el equinoccio vernal
Perihelio
Argumento del perihelio, 112°
Sol
Longitud del nodo ascendente, 58°
Dirección del equinoccio vernal
Posición de la Tierra en el perihelio del cometa, 9 de Febrero de 1986
Máxima proximidad el 27 de Noviembre de 1985
Inclinación orbital, 162°
Nodo ascendente
Trayectoria del cometa "por debajo" del plano de la eclíptica
Polvo meteórico abandonado por el cometa

8. Orbita de Halley con respecto al plano orbital terrestre



9. Halley y los planetas

porque el material expulsado por el núcleo se mueve rápidamente para formar dos colas. Ocasionalmente la coma exhibe cabelleras resplandecientes, las cuales acaso sean consecuencia de un calentamiento irregular de la superficie del núcleo.

A medida que se aproxima al Sol, el cometa comienza a generar también una cola de plasma y otra de polvo. La primera es un chorro recto de partículas cargadas apuntado casi exactamente en sentido opuesto al Sol. La segunda forma una curva cuya concavidad está dirigida hacia las posiciones anteriores del cometa. Ambas colas provienen del material expulsado de la superficie del núcleo, por lo que son más pronunciadas cuando el cometa está próximo al Sol.

Los átomos cargados eléctricamente y las moléculas del material expelido son obligados a entrar en la cola de plasma por acción del viento solar, que consiste en una corriente de protones y electrones que se mueven alejándose del Sol. La colisión entre el viento solar y los iones que manan del cometa deforman el campo magnético interplanetario, de tal manera que las líneas del campo magnético envuelven la cabeza del cometa y se prolongan a lo largo de una línea radial en dirección opuesta al Sol. Los iones del cometa se ven forzados a moverse en espiral alrededor de esas líneas radiales y son esos iones los que constituyen la cola de plasma.

Vemos esa cola porque los iones emiten luz. Esta es primordialmente azul a causa de las emisiones azules de las moléculas de monóxido de carbono ionizadas positivamente (CO^+). A menudo, la cola genera rizos y ondas provocadas por las irregularidades en la expulsión de materia del núcleo o en el mismo viento solar. A veces, la cola se separa del núcleo, tras lo cual se forma otra. Tales separaciones parece que ocurren cuando el cometa atraviesa alguna frontera dentro del campo magnético alrededor del Sol. La organización de las líneas de este campo se asemeja a un molinete compuesto de sectores espirales. Estos sectores son tales que, si en uno de ellos las líneas se dirigen al Sol, en el contiguo se alejan de él. Cuando un cometa cruza la frontera entre dos sectores, el cambio de sentido en las líneas de campo hace que se suelte la cola de plasma. Mientras el cometa viaja por el sector siguiente se forma una cola nueva.

La cola de polvo se forma con las partículas de polvo que desprende el núcleo. Estos gránulos, cada uno de un micrometro de diámetro aproximadamente, dispersan la luz solar con intensidad máxima en la zona amarilla del espectro. Así pues, la cola es amarilla.

Esta cola se genera porque el polvo de la coma es impelido a alejarse del Sol por la presión de la luz solar. (La luz posee cantidad de movimiento y ejerce una presión sobre toda superficie que ilumina.) Así, cada grano se ve atraído radialmente por el campo gravitatorio del Sol y repelido también radialmente por la presión de la luz solar. De esta manera, los granos de más de un micrometro de grosor se ven dominados por la atracción gravitatoria y acaban orbitando en torno al Sol formando un cinturón a lo largo de la trayectoria del cometa. Los granos más pequeños se ven dominados por la presión de la radiación y se alejan del Sol.

Estos granos más pequeños son los que forman la cola de polvo [*véase la figura 12*]. Aquí he supuesto que las partículas tienen el mismo tamaño, por lo que la luz actúa sobre todas ellas con la misma intensidad. Los gránulos, al soltarse, se mueven siguiendo una trayectoria curva que se aleja del Sol, ya que se separan del cometa animados de cierta cantidad de movimiento debida

Capa de polvo esponjoso, de 1 a 10 centímetros de espesor
Polvo expulsado por la presión del gas liberado
Posible núcleo rocoso
Hielo y polvo
Luz solar
Sublimación de hielo y liberación de gas

10. Un modelo de núcleo de cometa

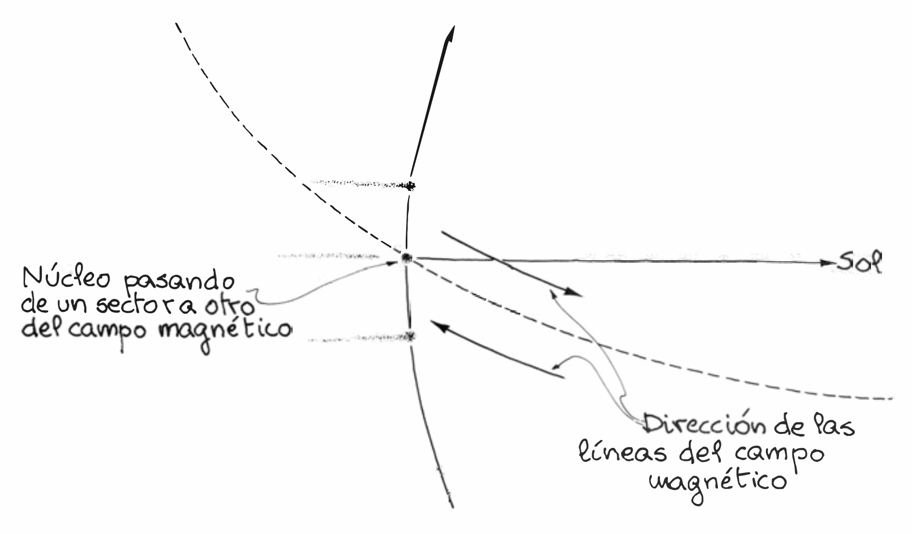
al movimiento mismo del cometa. En nuestro ejemplo, las partículas liberadas por el cometa al comienzo de su órbita se han desplazado hacia afuera para formar el extremo externo de la cola. Las partículas liberadas posteriormente no se han alejado tanto y forman la parte de la cola más cercana al núcleo. Este mecanismo provoca la ilusión de que las partículas de polvo expelidas del núcleo se mueven a lo largo de la cola.

En la realidad, las partículas que componen la cola son de tamaño variable. Las más pequeñas son impulsadas hacia afuera con mayor rapidez que las mayores, lo que hace que la cola sea más gruesa de lo que sería si las partículas fueran de tamaño uniforme. La forma de la cola puede alterarse aún más por erupciones de polvo en el núcleo.

Normalmente, la cola de polvo se hace más ancha y larga después del paso del cometa por su perihelio. Para entonces, el núcleo ya habrá dado la vuelta en torno al Sol, mientras que las partículas de polvo anteriormente desprendidas no. La longitud aparente de la cola depende de la visual del observador. Si la cola corre aproximadamente paralela a la visual, ocupará sólo un pequeño ángulo en el campo visual. La mejor perspectiva se presenta cuando la visual es perpendicular a la cola.

En ocasiones, un cometa puede presentar una anticola que parece apuntar al Sol. Se trata de un fenómeno ilusorio. La anticola es, en realidad, una vista casi de canto de las partículas mayores desparramadas a lo largo de la órbita del cometa. Estas son visibles si el observador está junto al plano orbital del cometa y si su visual es casi coincidente con la trayectoria descrita por el objeto. En tales condiciones, la visual atraviesa polvo suficiente para hacer visible la anticola.

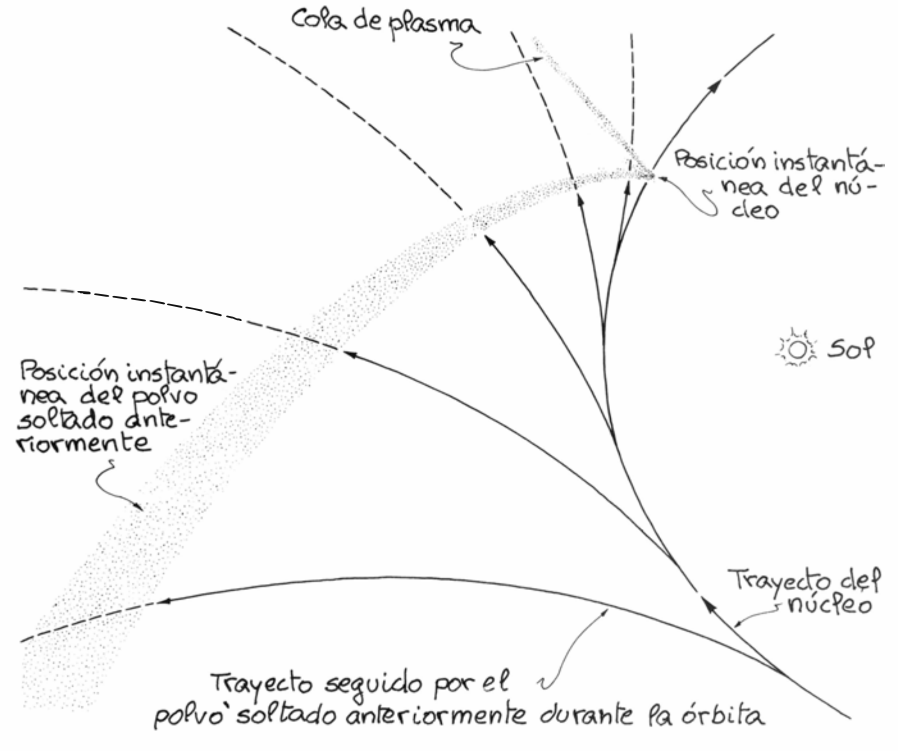
La cola de polvo del cometa Halley será probablemente corta, estrecha y recta antes de que el Sol la esconda. Si las condiciones de observación son buenas, se dejará ver extendiéndose casi verticalmente desde la cabeza del cometa. Inmediatamente antes del perihelio, la cola empezará a ensancharse, pero el resplandor del Sol impedirá esa observación. Más adelante, a finales de febrero, cuando el cometa salga del resplandor, su cola será más ancha, más larga y más pronunciada. Probablemente a principios de abril la visión de la cola sea óptima; apuntará generalmente al oeste. A mediados de abril la cola empezará a disminuir. La anticola se formará cuando el cometa reaparezca a finales de febrero y será débil y corta, por lo que sólo podrá verse en condiciones muy favorables.



11. Fuerzas sobre la cola de plasma

Las mayores partículas de polvo lanzadas por un cometa yacen en un cinturón a lo largo, más o menos, de la trayectoria orbital del cometa. Esas partículas prosiguen orbitando en torno al Sol. Si la Tierra atraviesa el polvo residual del cometa, las partículas arden por ablación cuando penetran en la atmósfera, dando lugar a lluvias de meteoritos. Nuestro planeta atraviesa la órbita del cometa Halley dos veces al año. Las lluvias de meteoritos Eta Aquarid, el pasado mayo, y la Orionid, en octubre, fueron ambas resultado de pasos anteriores del cometa Halley alrededor del Sol.

Se cree que el núcleo del cometa Halley tiene una masa de 2×10^{14} kilogramos; el diámetro medio del cometa es de unos 10 kilómetros. Las proporciones de polvo y materia sólida respecto al hielo son las medias de los demás cometas. En su último paso por el sistema solar interior perdió del orden de 2×10^{11} kg de materiales, lo que equivale a una capa superficial de unos dos metros de espesor. Pero, aun perdiendo materia a ese ritmo, al cometa Halley todavía le queda suficiente para dar muchas vueltas alrededor del Sol.



12. Cómo se forma la cola de polvo del cometa

Libros

Sobre Galois y Poisson, ecosistemas terrestres, controversias geológicas e inteligencia artificial

Javier de Lorenzo, Ramón Margalef, Salvador Reguant y Jordi Domingo

OBRA D'EVARISTE GALOIS, a cura d'Antoni Malet. Institut d'Estudis Catalans; Barcelona, 1984. Evariste Galois es una de las figuras más atractivas en la historia de la matemática. Por la brevedad de su vida –25 de octubre 1811 - 31 de mayo 1832: 20 años, 7 meses y 6 días–, por su actuación político-revolucionaria, por su obra... Su rechazo en el ingreso a la Escuela Politécnica ha quedado como sombra en las Grandes Escuelas francesas. Al presentarse Poincaré, el tribunal pensaba que, de aprobar, podría ser el número uno de su promoción, pero el problema era ese aprobado; Poincaré dibujaba muy mal y obtenía cero en dibujo, nota eliminatoria. La sombra de Galois hizo que el tribunal le puntuara con un uno en dibujo y Poincaré pudo ingresar...

Y, sin embargo, la obra de Galois, muy reducida en extensión, no ha sido publicada completa hasta 1962 en Francia, al cuidado de Bourgne y Azra. En España, que yo sepa, sólo el "Prefacio" fue editado en castellano, por mí, en 1971. Ahora Antoni Malet publica una selección de la obra de Galois en catalán. Tras una breve introducción, situando al personaje y su obra algebraica –pp. 5-12–, la antología comprende: *a.* "Prefacio" y "Discusiones sobre los progresos del Análisis puro" –pp. 14-20–; *b.* "Memoria sobre las condiciones de resolubilidad de las ecuaciones por radicales" –pp. 22-39–; *c.* Carta a Augusto Chevalier –pp. 51-56–; *d.* Además de las notas introductorias a cada apartado, unas notas finales de carácter aclaratorio a los escritos matemáticos y un apéndice con el Acta de la sesión de la Academia de Ciencias de París del 4 de julio de 1831, en la cual Lacroix y Poisson, realmente este último, informan desfavorablemente sobre la "Memoria" presentada por Galois a la Academia y que ésta, a la vista del informe, rechaza.

En los escritos de Galois cabe distinguir dos facetas: su aspecto humano y el contenido de la obra matemática. Del aspecto humano resalta un pathos trágico que culmina en una muerte ridícula, en duelo con dos patriotas republicanos por una "puta". De la obra matemática, su contribución a la teoría de grupos a través de la resolución de las ecuaciones algebraicas, aunque no empleara el término grupo. Ambas facetas han sido tratadas en otros lugares y en este breve librito Malet hace una exposición suficiente de ellas.

Voy a detenerme en un aspecto diferente: la incomprensión de la obra de Galois por sus contemporáneos, el rechazo de la "Memoria" por la Academia de Ciencias motivado por el informe de Poisson. A Malet le resulta difícil explicar que un hombre de la competencia de Poisson no haya visto la importancia del trabajo que tenía que informar, a pesar de atenuantes como, por una parte, reconocer que el escrito de Galois era muy conciso y difícil de leer; por otra, que Poisson estaba ligado a cuestiones de física, de matemática aplicada. Como última razón, Malet indica la crítica que Poisson hace a Galois: la memoria estaba dedicada a la resolución de las ecuaciones por radicales pero, en ella, el método indicado se mostraba realmente impracticable. Algo ya previsto por el mismo Galois: si se enfoca el problema de la resolución desde un punto de vista práctico, sus teorías son inútiles, textualmente: "En una palabra, los cálculos son impracticables." En cuanto a la concisión y dificultad del escrito de Galois, el propio Malet se hace testigo: sólo tras los trabajos de Hermite, Serret, Bertrand... y, fundamentalmente, tras la obra de Jordan de 1871, dicha "Memoria" se hará legible. Incluso las notas aclaratorias que Malet agrega a la Memoria, siguiendo básicamente a Jordan, muestran una extensión casi mayor que el trabajo original de Galois de seis folios. Con el reconocimiento de que el tema central, la noción de grupo que forman las raíces, no queda definido por Galois en momento alguno; más aún, sólo hacia finales de siglo se llega a su formulación, digamos, completa.

Como consecuencia, Malet agrega que los trabajos de Abel, de Galois, lo que hacen es iniciar una nueva perspectiva para el estudio de las ecuaciones: no importa la solución práctica, algorítmica, de las mismas; lo que importa es un enfoque estructural, las condiciones que han de cumplir raíces y coeficientes para que pueda hablarse de resolución.

Y es la razón de la incomprensión. No sólo el estilo o la preocupación de orden pragmático de Poisson y de algunos matemáticos de primeros del siglo XIX. Lo que se juega es otro tipo de hacer matemático. Que, en cuanto a contenido, se liga a las funciones elípticas –tema que aparece precisamente hacia 1831–; sin olvidar los nuevos enfoques de la geometría proyectiva, la diferencial... Pero es más que el contenido: se trata de una inversión conceptual por la que los cálculos y desarrollos algorítmicos quedan marginados, es el "saltar de puntillas sobre los cálculos" de Galois. Es lo que he denominado en otros lugares una ruptura epistemológica que decantan, precisamente, matemáticos como Galois, como Jacobi, como el principal iniciador, Abel, de quien esa ruptura podría tomar el lema: buscar la razón. Desde este punto de vista la obra de estos matemáticos supone en terminología que no me convence, pero de moda tras los trabajos de Kuhn, de Lakatos, un cambio de paradigma y el surgimiento de una matemática revolucionaria, nueva, frente a una matemática que se convierte en clásica y, por ello, anticuada. Y aunque Malet no insista lo suficiente –ni, por supuesto, emplee los términos aquí mencionados–, no podía escapársele este hecho de ruptura o cambio cualitativo que sólo podrá imponerse con lentitud y que explica la incomprensión de un Poisson respecto a la nueva matemática. En cualquier caso, es interesante que, al menos, parte de la obra de Galois haya sido traducida al catalán, dando fe de ese cambio cualitativo en el hacer matemático. (J.L.).

ECOLOGICAL SYSTEMS OF THE GEOBISOSPHERE. 1. ECOLOGICAL PRINCIPLES IN GLOBAL PERSPECTIVE, por Heinrich Walter y Siegmar-W. Breckle. Springer-Verlag; Berlin, 1985. Es el primer volumen de una obra de síntesis ecológica, que se centra y limita al estudio de la vegetación terrestre. Este volumen trata de los principios ecológicos generales, el segundo se ha de ocupar de las zonas tropical y subtropicales, y el tercero de las zonas templadas y polares. Este volumen es la

traducción del original alemán que apareció en 1983. Pocas personas habrá que tengan más experiencia en los ecosistemas terrestres de todo el mundo que el profesor Walter, por lo que es de esperar una información copiosa, sensata y de carácter naturalista. Esto se aprecia mucho hoy día, en que tantos textos ecológicos derivan quizá demasiado hacia la física, la química y, no digamos, hacia los modelos matemáticos.

Una parte de estos materiales se habían incluido en obras anteriores de H. Walter, como es su *Vegetation der Erde*, pero aquí están resumidos, reorganizados y con las adiciones pertinentes. Los temas centrales sobre los que los autores organizan su texto son: *1*) La zonación ecológica evidente en la superficie de los continentes, *2*) Los ciclos en los ecosistemas de distinta organización bajo la acción del fuego y de los animales, *3*) Factores limitantes, básicamente temperatura y disponibilidad de agua, *4*) Producción primaria y factores y agentes que la afectan, *5*) Competencia, incluyendo el nivel de las raíces y *8*) Sucesión y clímax. En un interesante capítulo se presenta un ejemplo de mosaico de vegetación en los trópicos venezolanos.

En el libro está presente la mejor tradición europea del conocimiento de la flora y de la vegetación; pero persisten con ella los problemas de definir unidades, generalmente bien resueltos en el sentido de no considerarlas más que como comunidades de trabajo. Este punto de vista se refleja también en la bibliografía, donde no se incluye nada de Odum ni de Tansley y muy poco de Whittaker. Animales y bacterias quedan relegados a segundo término o se ven como una complicación innecesaria; pero los autores hacen un esfuerzo por comparar los ecosistemas continentales con los oceánicos, muy diferentes en el valor de la relación producción primaria/fitomasa (4200 veces mayor en los ecosistemas acuáticos) y reconocen que la investigación de los ecosistemas empezó en la limnología. (R. M.)

GRANDES CONTROVERSIAS GEOLOGICAS, por Anthony Hallam. Editorial Labor, S.A.; Barcelona, 1985. Este libro presenta un interés excepcional por el contenido y el enfoque dado a su tratamiento, aparte de la reconocida personalidad del autor, catedrático de geología de la Universidad de Birmingham.

Se trata, en realidad, de la historia de los conceptos fundamentales y de la evolución de las teorías en el campo de la geología a lo largo de los dos últimos siglos, justamente las centurias en que la geología adquirió su definitivo *status* de ciencia. El libro que analizamos reconstruye esta historia a través del análisis de las grandes controversias. Cómo se produjeron, cuáles fueron sus epígonos y cuáles sus condiciones ambientales y de formación científica. Es una historia de la geología moderna a través de la evolución y vicisitudes de sus teorías más generales o más polémicas.

Sus grandes capítulos, indicados como subtítulo en la propia portada de la obra, nos dan una idea clara del contenido y del orden de tratamiento, que es cronológico, con solapamientos y recurrencias: (*1*) Neptunistas, vulcanistas y plutonistas; (*2*) Catastrofistas y uniformitaristas; (*3*) La era glacial; (*4*) La edad de la Tierra y (*5*) La deriva de los continentes.

Algunos de estos temas se refieren a problemas globales, desde distintos puntos de vista, como es el caso del neptunismo de Werner o de la actual teoría de la tectónica de placas; aunque, obvio es decirlo, con un sentido de globalidad muy distinto en uno y otro caso. El sentido de globalidad en el neptunismo se refiere a un único modo de interpretar la formación de las rocas, elevando a teoría general unos hechos concretos supuestamente probados, simplificando el conjunto de procesos petrogenéticos a favor de una explicación casi única. En la teoría de la tectónica de placas, los datos obtenidos en el estudio de la situación y características de la litosfera se hacen servir para confeccionar una explicación coherente con los obtenidos en el estudio de los diversos procesos y fenómenos geológicos. El libro de Hallam permite contemplar un progreso colectivo de maduración en el conjunto de los geólogos, cada vez hacia posiciones más acordes con el espíritu crítico propio de la ciencia.

Si atendemos a este aspecto de "cientificidad progresiva" de la geología, se ha producido y continúa produciéndose una controversia de fondo entre los partidarios de considerar más "científica" una aproximación a la física, o un uso creciente de sus métodos y perspectivas, y los que, sin ser contrarios al uso constante e inevitable de conceptos y métodos físicos en geología, no aceptan que la opinión de los físicos pueda condicionar de una manera absoluta las teorías geológicas.

Este trasfondo condicionó la gran controversia sobre la edad de la Tierra. La falta de métodos adecuados de datación permitió la incursión de físicos como Lord Kelvin, que intentaron hacer cálculos de la edad global de nuestro planeta y del Sol a partir de datos e hipótesis astronómicas, con un desconocimiento casi absoluto de los datos geológicos, e incluso con cierto desprecio para la labor callada y constante de los geólogos regionales y de las propias teorías geológicas generales. Los resultados reales fueron funestos, forzando a muchos geólogos a interpretaciones no queridas o desacordes con los datos conseguidos para adaptarse al condicionamiento absoluto de la edad global impuesto por las conclusiones de los físicos. En realidad, estos efectos perniciosos se prolongaron hasta el descubrimiento de la radiactividad natural. Con ello cayó el fundamento de las teorías de Lord Kelvin referente a la evolución térmica de la Tierra, por una parte y, por otra, se crearon los fundamentos de la radiometría aplicada a la geología, es decir, se inició la geocronometría con bases fidedignas de la datación de los materiales geológicos.

Del libro de Hallam se deduce cómo el avance de la geología depende del progreso científico-tecnológico general, aunque también pueda aceptarse que la actitud y métodos de cada ciencia son lo suficientemente peculiares y específicos para que sean los cultivadores de cada una de ellas los que tengan la última palabra al enjuiciar las distintas hipótesis y teorías de su propia disciplina. La evolución de las ideas sobre la edad de la Tierra es un claro ejemplo; y cuando se ha seguido otro camino, los resultados, lamentables, se han encargado de mostrar el fracaso. Las explicaciones de la historia de la Tierra hubieran sido más correctas de haber atendido a paleontólogos (Darwin) y estratígrafos (Lyell y otros) que no a físicos (Kelvin).

En la permanencia de las leyes físicas y la posible variación de sus parámetros en diferentes momentos de la historia de la Tierra se centra la vieja polémica catastrofismo-uniformitarismo que, nacida a fines del siglo XVIII, perdura aún en el presente, si bien con un planteamiento radicalmente distinto.

En sus primeros tiempos, la controversia se integraba en el debate general en torno a una cosmovisión religiosa (pretendidamente basada en la Biblia) o una cosmovisión laica. Antagonismo que más tarde se abandonó con la salvedad excepcional del brote creacionista norteamericano de los últimos años, sin incidencia relevante en Europa. Pero la controversia continúa y

con frecuencia se mueve en planos incorrectos o carente de claridad en su planteamiento. Existe confusión terminológica y también cierta incapacidad por parte de algunos en captar los diversos órdenes de magnitud espaciotemporales, por lo que se producen discusiones estériles en torno a la manera como se han producido los distintos fenómenos geológicos. El triunfo temprano del uniformitarismo constituyó la baza fundamental para dotar de soporte "científico" a la geología. La controversia ha experimentado vicisitudes de menor interés y aun inútiles y perniciosas para el desarrollo de la disciplina. Un cierto relativismo y moderación va permitiendo aceptar a la vez el catastrofismo y el uniformitarismo o, si se quiere, un catastrofismo parcial dentro de un uniformitarismo sustantivo.

En todas las grandes controversias subyace la suma de observaciones realizadas sobre el terreno; los viajes conjuntos para observar diversos fenómenos en el campo forman parte del aspecto humano en las discusiones más o menos civilizadas entre los diversos epígonos de las hipótesis y teorías contrapuestas. El hecho de que la observación de las mismas rocas, montañas y fenómenos geológicos en excursiones conjuntas posibilite interpretaciones distintas, y aun opuestas, nos hace ver el peso de los prejuicios, de la propia formación intelectual y de la diversa selección que el observador hace de lo que debe ser observado. Recuérdese a este propósito la disputa sobre la era glacial o la edad del hielo: se discrepa porque los puntos de referencia de los autores son distintos en su ubicación, área o región de trabajo. Sin olvidar que la inercia intelectual y el conservadurismo son lastres pesados que debilitan mucho la vista del científico, cuando la observación de determinados hechos o fenómenos requiere una explicación nueva o insospechable antes de esta misma observación. Asimismo, una nueva interpretación usada sin criterio, o con pretensiones excesivamente generalizadoras, provoca el rechazo de otros más exigentes.

Hallam concluye su libro con un interesante capítulo en el que reflexiona sobre la historia relatada por él mismo, con puntos de vista agudos y sugerentes acerca de la evolución de la ciencia geológica y con una discusión crítica acerca de cómo esta evolución sigue las pautas o modelos que para la evolución o historia de la ciencia han elaborado los filósofos de la ciencia. Se refiere en particular a Popper, Kuhn y Lakatos mostrando una predilección por este último. Y adelanta una reflexión sobre el futuro de la teoría de la tectónica de placas desde el punto de vista del movimiento pendular que se produce entre los dos polos de cualquier controversia científica profunda. (S. R.)

INTELIGENCIA ARTIFICIAL: CONCEPTOS Y PROGRAMAS, por Tim Hartnell. Ediciones Anaya Multimedia, S.A., 1985. En los últimos años la oferta bibliográfica en torno al mundo de la informática ha experimentado un crecimiento singular, desde libros especializados para profesionales hasta revistas dirigidas a los aficionados. Durante 1985 varias editoriales coincidieron en sacar a luz diversas colecciones de índole varia, acomodadas a veces a distintos ordenadores puestos a la venta.

Una de las aplicaciones de los ordenadores destaca claramente de entre las demás porque abre las puertas a una nueva concepción en la utilización de las máquinas. Me refiero a la "inteligencia artificial". Tema este explotado, además, por la literatura y el cine, dos fuentes que alimentan nuestra imaginación y nuestros temores. Pero, ¿puede una máquina pensar realmente? ¿Cuál es la verdadera naturaleza de la inteligencia? ¿Se llegará alguna vez a construir una máquina que pueda participar de esa inteligencia? Las respuestas, o mejor dicho, lo que Tim Hartnell piensa sobre estas cuestiones se encuentra entre los comentarios dedicados a cada tema y a cada uno de los programas.

La orientación del libro es fundamentalmente práctica. Expone una serie de conceptos, experiencias y programas para que el lector pueda sacar sus propias conclusiones. La introducción termina con una sugestiva proposición: "Espero que este libro transmita al menos algo de la fascinación que yo experimenté, y que tú, lector, sientas la misma emoción que yo sentí al ejecutar los programas."

En la primera sección del libro, bajo el título genérico de "el pensamiento", expone qué es lo que puede hacer y lo que no puede hacer una máquina, es decir, hasta dónde puede llegar la inteligencia artificial. "Es cierto que los programas incluidos en este libro permiten al ordenador dar muestras de inteligencia respondiendo a situaciones, tomando decisiones y actuando de acuerdo con ellas. Pero no se sugiere en ningún momento que el ordenador sea consciente de sus actos. ¿Estamos, pues, justificados cuando afirmamos que estamos produciendo 'inteligencia artificial'? Yo creo que sin el tipo de percepción que reconoce cosas tales como el 'efectismo' de un poema o la incongruencia de una respuesta, no podemos sugerir que nos encontremos ante ningún tipo de inteligencia". Según se deduce de ello, el autor considera la percepción y la conciencia de sus actos como prerrequisito para afirmar que un sistema es inteligente.

Más adelante concreta algunas de las manifestaciones de un comportamiento inteligente. Dos son particularmente prometedoras: la capacidad de realizar inferencias lógicas y la capacidad de almacenar resultados para ser luego utilizados a modo de experiencia. Los términos empleados en el libro para designar estos dos procesos son bastante ambiguos y conducentes a confusión (el autor habla de "razonamiento" y "aprendizaje", exclusivos de la actividad humana).

La segunda sección del libro se dedica a lo que el autor denomina "investigación", aunque los términos exploración o búsqueda sean probablemente más exactos. En estos capítulos presenta varias maneras de organizar la información, para poder luego examinar todas las opciones, evaluar distintas posibilidades y escoger entre opciones alternativas.

Los estudios en estas materias se iniciaron en el campo de los juegos: matemáticos, de cartas (póker), backgammon, othello, dominó, shogi, go-moku y ajedrez. Las ventajas principales que presentan los juegos, en contraste con los problemas reales, es que se rigen mediante un conjunto de reglas fijo (las reglas del juego), su universo es reducido (las posibilidades están acotadas a un número relativamente pequeño) y los criterios acerca del resultado obtenido son claros (gana la partida o la pierde). El autor se entretiene en una versión simplificada del juego de las damas.

En la tercera sección aborda el tema del habla, la comprensión del lenguaje natural, los problemas derivados del análisis sintáctico y semántico y las restricciones a imponer.

El programa que elige para ilustrar este tema es realmente significativo. Se trata de un programa realizado por Joseph Weizenbaum, profesor de informática del Instituto de Tecnología de Massachusetts, entre los años 1964 y 1966. "Eliza" es el nombre del programa; su objetivo: parodiar un psiquiatra rogeriano, cuyo método de trabajo consiste en dejar que sea siempre el paciente quien dirija la conversación. A raíz de la reacción de la gente ante el programa, una vez éste se hizo público, Weizenbaum realizó un estudio sobre la inteligencia artificial y las relaciones y dependencias que se de-

sarrollan entre el hombre y la máquina, publicado en 1976 con el titulo "Computer Power and Human Reason". En este trabajo Weizenbaum dice que no había calculado el "poder de engaño" que un programa, simple por otro lado, podía crear en la gente normal.

Aunque no es este el lugar para comentar las consideraciones del autor de Eliza, sí merece la pena recoger alguna de sus conclusiones para poder tener una idea del alcance de este tipo de aplicaciones. Weizenbaum considera que hay ciertas áreas a las que nunca se debería permitir que accedieran las máquinas, aun cuando éstas estuvieran capacitadas para hacerlo. Esta tajante afirmación no es compartida por Tim Hartnell, ni por otros investigadores ni médicos psiquiatras que experimentaron con Eliza. Para ellos, "¿acaso no debe explorar el terapeuta todos los medios a su alcance, aunque sólo sea para comprobar si alguno de ellos resulta ser genuinamente efectivo?" Se descubre aquí otro punto fundamental en el campo de la aplicación práctica de la inteligencia artificial. Cuando estas realizaciones interaccionan con sujetos no se ha de olvidar que la dignidad de la persona y sus derechos están por encima de las experiencias científicas, y en cada caso se deberá evaluar el fondo, la forma, los medios y las posibles repercusiones de este tipo de aplicaciones.

En lo que se refiere al programa Doctor presentado en el libro, reviste especial interés ver cómo un programa relativamente sencillo puede generar un diálogo efectista. No se han de olvidar, por otra parte, las respuestas incongruentes y fuera de contexto que aparecen de cuando en cuando.

El tema de la última sección es la ayuda. Se concreta en los denominados "sistemas expertos". Es éste el campo de la inteligencia artificial en el que se han dado los avances más significativos. Aquí, además, nadie se preocupa de si la máquina piensa o deja de pensar cuando el sistema muestra su capacidad o pericia en un determinado tema. Generalmente, se trata de problemas que requieren un profundo conocimiento de la materia en cuestión; y la misión de estos sistemas es la de poner al alcance de muchos el saber y la experiencia de los expertos. El problema principal estriba en llegar a conocer cuáles son los mecanismos que sigue el razonamiento de un experto a partir de las respuestas obtenidas para ir acotando el problema con preguntas cada vez más precisas hasta llegar a proponer una solución. Este proceso de extracción de la ciencia y la experiencia de una persona experta es la fase más delicada y costosa. Una vez especificados los conceptos y tipificada la manera de utilizarlos se traduce esta información y se introduce en el sistema con resultados más o menos afortunados.

Ilustra esta sección un conjunto de programas que son capaces de aprender, en el sentido de que son capaces de no volver a equivocarse en otra situación idéntica. El método consiste en indicar al final si la conclusión a la que se ha llegado es correcta o no. Estos programas son un buen ejemplo para comprender cómo un sistema puede ir almacenando información a través de la experiencia.

Finalmente, en lo que se refiere a la obra en conjunto, cabe destacar que la estructura del libro es muy didáctica. Cada sección introduce un tema con una exposición general. Luego se concreta en un programa, el cual, una vez expuesta la idea global del mismo, se desglosa e ilustra con ejemplos. Cada capítulo termina con un listado completo del programa en Basic, lo que permite experimentar con él. (J. D.)

Bibliografía

Los lectores interesados en una mayor profundización de los temas expuestos pueden consultar los trabajos siguientes:

TERAPEUTICA Y LUCHA CONTRA EL CANCER

CANCER MORTALITY IN THE UNITED STATES: 1950-1977. Frank W. McKay, Margot R. Hanson y Robert W. Miller. Monografía n.º 59 del Instituto Nacional del Cáncer, del Departamento para la Salud Pública de los Estados Unidos. Servicio de Publicaciones del Gobierno de EE.UU., 1982.

MASS SCREENING FOR CONTROL OF BREAST CANCER. Philip Strax en *Cancer,* vol. 53, n.º 3, págs. 665-670; 1 de febrero de 1984.

DECLINE IN U.S. CHILDHOOD CANCER MORTALITY, 1950 THROUGH 1980. Robert W. Miller y Frank W. McKay en *Journal of the American Medical Association,* volumen 251, número 12, páginas 1567-1570; 23/30 de marzo de 1984.

RAYOS COSMICOS DE CISNE X-3

AN ASTRONOMICAL PUZZLE CALLED CYGNUS X-3. R. M. Hjellming en *Science,* vol. 182, n.º 4117, págs. 1089-1095; 14 de diciembre de 1973.

GAMMA-RAY ASTRONOMY. Dirigido por H. S. W. Massey, R. D. Wills and A. W. Wolfendale. Royal Society of London, 1981.

HIGH ENERGY ASTROPHYSICS. Malcolm S. Longair. Cambridge University Press, 1981.

LA SEÑAL DEL CALCIO

CA^{++} UPTAKE BY RAT KIDNEY MITOCHONDRIA AND ITS DEPENDENCE ON RESPIRATION AND PHOSPHORYLATION. Frank D. Vasington y Jerome V. Murphy en *The Journal of Biological Chemistry,* vol. 237, n.º 8, págs. 2670-2677; agosto, 1962.

ATP-DEPENDENT CA^{2+}-EXTRUSION FROM HUMAN RED CELLS. H. J. Schatzmann en *Experientia*, vol. 22, n.º 6, págs. 364-365; 15 de julio de 1966.

PURIFICATION AND PROPERTIES OF AN ADENOSINE TRIPHOSPHATASE FROM SARCOPLASMIC RETICULUM. David H. MacLennan en *The Journal of Biological Chemistry,* vol. 245, n.º 17, págs. 4508-4518; 10 de septiembre de 1970.

PURIFICATION OF THE (CA^{2+}-MG^{2+})-ATPASE FROM HUMAN ERYTHROCYTE MEMBRANES USING A CALMODULIN AFFINITY COLUMN. Verena Niggli, John T. Penniston y Ernesto Carafoli en *The Journal of Biological Chemistry*, vol. 254, n.º 20, págs. 9955-9958; 25 de octubre de 1979.

RETAZOS LITOSFERICOS

CORDILLERAN SUSPECT TERRANES. Peter J. Coney, David L. Jones y James W. H. Monger en *Nature*, vol. 288, n.º 5789, págs. 329-333; 1980.

SUSPECT TERRANES AND ACCRETIONARY HISTORY OF THE APPALACHIAN OROGEN. Harold Williams y Robert D. Hatcher, Jr., en *Geology*, vol. 10, n.º 10, págs. 530-536; octubre, 1982.

EL CRECIMIENTO DE NORTEAMERICA. David L. Jones, Allan Cox, Peter Coney y Myrl Beck en *Investigación y Ciencia*, n.º 76, págs. 30-45; enero, 1983.

PHANEROZOIC ADDITION RATES TO THE CONTINENTAL CRUST AND CRUSTAL GROWTH. A. Reymer y G. Schubert en *Tectonics*, vol. 3, n.º 1, págs. 63-77; febrero, 1984.

RESPIRACION CUTANEA EN LOS VERTEBRADOS

THE EVOLUTION OF AIR BREATHING IN VERTEBRATES. David J. Randall, Warren W. Burggren, Anthony P. Farell y M. Stephen Haswell. Cambridge University Press, 1981.

A MODEL FOR EVALUATING DIFFUSION LIMITATION IN GAS-EXCHANGE ORGANS OF VERTEBRATES. J. Piiper en *A Companion to Animal Physiology*, dirigido por C. R. Taylor, K. Johansen y L. Bolis. Cambridge University Press, 1982.

Cutaneous Gas Exchange in Vertebrates: Design, Patterns, Control and Implications. Martin E. Feder y Warren W. Burggren en *Biological Reviews*, vol. 60, págs. 1-45; 1985.

VUELO DE PROPULSION HUMANA

FLUG DURCH MUSKELKFRAFT. Hans-Georg Schulze y Willi Stiasny. Fritz Kanpp, Frankfurt, 1936.

GOSSAMER ODYSSEY: THE TRIUMPH OF HUMAN-POWERED FLIGHT. Morton Grosser. Houghton Mifflin Company, 1981.

BICYCLING SCIENCE. Frank Rowland Whitt y David Gordon Wilson. The MIT Press, 1982.

TARJETAS INTELIGENTES

SMART CREDIT CARDS: THE ANSWER TO CASHLESS SHOPPING. Stephen B. Weinstein en *IEEE Spectrum*, vol. 21, n.º 2, págs. 43-49; febrero, 1984.

SMART CARDS-THE ULTIMATE PERSONAL COMPUTER. Jerome Svigals. Macmillan Publishing Company, 1985.

COMPORTAMIENTO DE LIQUIDOS EN INGRAVIDEZ

AN INTRODUCTION TO FLUID DYNAMICS. G. K. Batchelor. Cambridge University Press; London, 1967.

THE BREAKING OF AXISYMMETRIC SLENDER LIQUID BRIDGES. J. Meseguer en *Journal of Fluid Mechanics*, vol. 130; 1983.

LONG LIQUID BRIDGES. I. Da Riva e I. Martínez en *EUROSPACE, Proceedings of the Symposium on Industrial Activity in Space*, Stresa, Italia; 2/4 de mayo de 1984.

RESULTS OF SPACELAB-1. 5th European Symposium on Material Sciences under Microgravity. ESA-SP-222, ESA; París, 1985.

NUMERICAL AND EXPERIMENTAL STUDY OF THE DYNAMICS OF AXISYMMETRIC SLENDER LIQUID BRIDGES. J. Meseguer y A. Sanz en *Journal of Fluid Mechanics*, vol. 153; 1985.

JUEGOS DE ORDENADOR

ARTIFICIAL INTELLIGENCE THROUGH SIMULATED EVOLUTION. Lawrence J. Fogel, Alvin J. Owens y Michael J. Walsh. John Wiley, 1966.

ADAPTATION IN NATURAL AND ARTIFICIAL SYSTEMS. John H. Holland; University of Michigan Press, 1975.

GENETIC ALGORITHMS FOR THE TRAVELING SALESMAN PROBLEM. John Grefenstette, Rajeev Gopal, Brian Rosmaita y Dirk Van Gucht en *Proceedings of an International Conference on Genetic Algorithms and Their Applications*, dirigido por John J. Grefenstette. The Robotics Institute, Carnegie-Mellon University, 1985.

TALLER Y LABORATORIO

INTRODUCTION TO COMETS. John C. Brandt y Robert D. Chapman. Cambridge University Press, 1981.

COMETS. David W. Hughes en *Contemporary Physics*, vol. 23, n.º 3, págs. 257-283; mayo/junio, 1982.